

Learning from Chernobyl

As the accident that blackened the name of nuclear power fades from memory, openings present themselves for the technology to edge its way back into public favour.

The image is a resonant one and rests indelibly in the mind. The plant sits, fuming restlessly, while all around an inept Soviet bureaucracy crumbles into the ensuing chaos. No one knows how many will have died as a result of the radioactive cloud expelled by Chernobyl's number 4 reactor on 26 April 1986; the fact that people are still debating it (see pages 982 and 993) says enough.

Skip forward 20 years, and nuclear power is edging back into vogue. It wasn't just Chernobyl that drove it out of favour, of course: the Three Mile Island incident in Pennsylvania in 1979, a catalogue of economic and technical setbacks in several nations, and the surprising resilience of fossil fuels as a cheap and available source of energy had already seen to that.

As memories of these mishaps recede, other factors have arisen to bring nuclear power back into play. Energy prices are high again, and governments are seeking to tackle climate change by limiting fossil-fuel emissions. Economists from California to Calcutta are looking at pie-charts of their future energy supply and saying that nuclear power needs to play a role. Will it?

The answer is a qualified yes — provided that governments absorb the true lesson of Chernobyl. This is not that nuclear power is unsafe, but that it is unsafe in the hands of a corrupt, unaccountable, irresponsible political system that fails to take reasonable measures to protect its citizens. The future of nuclear energy does not hinge primarily on the development of a safer reactor or a more geologically reliable waste repository, but on the ability of states to build public trust in their ability to safely implement and manage the technology.

Building trust

This trust can be achieved in different ways. In France, where the public appreciates the centralized technocracy that brought it high-speed trains, Concorde and an independent nuclear deterrent, nuclear power is ubiquitous and widely accepted. Scandinavian nations have a different political tradition, in which inclusive decision-making may soon open the way for the world's first permanent nuclear-waste repositories. Elsewhere, however, the future of nuclear energy is uncertain. The mechanisms that will assure its acceptance are not yet in place.

The key elements of this equation are the same the world over: nuclear power's real and perceived links with nuclear weapons; the available technology for power generation; its safety and economics; and options for clean-up and waste disposal.

Recent events in Iran serve as a painful reminder of the interaction between nuclear power and nuclear weapons. From the days of the 'Atoms for Peace' movement in the 1950s, advocates of the former have sought to separate the two, but they are inextricably linked. Public acceptance of nuclear power in Europe, Japan and the United States would benefit from a credible strategy to contain the proliferation of nuclear weapons. Expansion of nuclear power elsewhere,

on the back of further proliferation of nuclear weapons, could have disastrous consequences.

The technology of nuclear power plants continues to improve (see *Nature* **429**, 238–240; 2004). Chernobyl was a vastly archaic reactor whose safety systems, regulation and management were not close to acceptable standards. Modern working reactors are not susceptible to Chernobyl-style accidents, and some designs now under consideration could be safer still.

But safe reactor design, unfortunately, provides little protection against current fears — floated again this week in a sceptical report from a committee of British members of parliament — that nuclear plants may be vulnerable to terrorist attack.

Counting the cost

The economics of such plants are subject to fierce debate (see page 984). The deregulated power-generation markets that have taken shape over the past two decades have little appetite for nuclear power's combination of high build costs, low running costs and uncertain future liabilities. Old debates about how many cents it costs to produce electricity have been superseded by a more subjective discussion about what goes in the bucket labelled 'costs'. Restoring sites to a pristine state may be prohibitively expensive, as may the permanent disposal of waste and spent fuel. The construction of new plants will require either financial guarantees from the state (in Britain or the United States) or direct government involvement (in India or China).

Finally, nuclear-waste disposal remains the industry's Achilles' heel. For governments that advocate nuclear power to offer no solution — and leave spent fuel and other waste on the surface for future generations to deal with — is an abdication of responsibility. The threat of terrorist attack on nuclear power stations, as well as the risk that their spent fuel could be stolen and used for such an attack, renders the notion of long-term, localized waste storage at multiple sites even less tenable than it was before.

Different approaches are being taken to the management of waste disposal. Yucca Mountain in Nevada is in grave danger of becoming an expensive monument to failure (see page 987). The site was selected by default when Nevada was too weak to remove itself from the process — a hopeless, unscientific approach that may now reap what it sowed. Scandinavian states are doing a little better; Finland is winning support for a repository on the basis of a continuing nuclear energy programme, and Sweden is doing so on a promise to wrap the whole thing up. France has made some headway in site selection and can be relied on to address the issue with its customary

"Restoring sites to a pristine state may be prohibitively expensive, as may the permanent disposal of waste and spent fuel."



determination. Britain has to start again from scratch, and is using its Committee on Radioactive Waste Management as an interesting, if not entirely convincing, experiment in public consultation.

So far, India and China, the biggest likely builders of nuclear power stations in the next 20 years, don't have much to say about waste disposal. Time will tell if either of them can handle the issue in an environmentally responsible way. However, if national pride in nuclear technology is a significant factor, the French example suggests that nuclear power has a solid future in Asia, with or without a waste repository.

In the West, however, the future options for nuclear power are far narrower than the heat of the current debate would suggest. Abandonment, as embraced fleetingly by the previous German government, isn't going to happen. The kind of major build-up envisaged

before Three Mile Island and Chernobyl (see *Nature* **244**, 392; 1973 and *Nature* **257**, 346; 1975) isn't coming either.

Instead, nations are likely to tread a path somewhere between replacing some existing nuclear power capacity and its mild augmentation. Given global warming, high energy costs and doubts about the reliability of the oil supply, the latter approach has much to commend it, although it should not be pursued at the expense of renewable energy.

Nuclear energy's technical elegance has always appealed to the hearts and minds of scientists and engineers, who have been unusually prominent among its public advocates for half a century. Throughout, these advocates have promised to present to the public a clean and complete nuclear fuel cycle. Now it is time to stand and deliver. ■

Drugs tests on trial

Britain's clinical-trial regulator has no good options.

Following an alarming episode in London last month, in which six drug-trial participants needed emergency treatment, the UK Medicines and Healthcare Products Regulatory Agency (MHRA) says it will change the way it regulates clinical trials, at least temporarily. But this may prove more easily said than done.

In the trial on 13 March, six healthy subjects suffered violent reactions within minutes of ingesting an antibody drug candidate, TGN1412, which was being developed to treat autoimmune diseases such as rheumatoid arthritis. Initial investigations suggest that the antibody itself was responsible for the side effects (see *Nature* **440**, 855–856; 2006). On 5 April, the MHRA said it will seek advice from outside experts in determining whether drug candidates with novel modes of action should be allowed to enter clinical trials.

The incident at London's Northwick Park Hospital has drawn attention to the limitations of preclinical animal trials in determining the safety of drugs in humans, especially for 'humanized' antibody drugs that are targeted at mimicking human biological processes. It has also sparked some debate about whether the participants were sufficiently aware of the dangers they faced.

For the regulator, the immediate question is whether the existing rules strike the right balance between safeguarding trial participants

and promoting the study of potentially valuable cures. Previously, the MHRA allowed initial, small-scale human safety trials to go ahead on the basis of successful animal trials and a description of how the compound works.

Now the agency says it will allow such trials to proceed only after review by a panel of outside experts. However, companies that have drug candidates up their sleeves don't want information on them to be shared, and any outside panel worth its salt is bound to contain people who work with rival companies. So such a provision could lead drug developers to turn their backs on Britain as a location for early-stage clinical trials.

The best approach is probably that practised by the US Food and Drug Administration (FDA), the only drug regulator in the world with the in-house expertise to conduct such reviews by itself in strict confidence. The FDA, which is partly supported by fees levied on drug-makers eager to enter the lucrative US market, has 9,000 staff compared with the MHRA's 800 (although the FDA does handle food as well as drug safety).

One theoretical option would be a Europe-wide body set up to regulate and approve clinical trials, but the practical problems of constructing and operating such an agency would be daunting. In the interim, the MHRA may struggle to perform additional screening while satisfying confidentiality requirements. ■

"The incident has drawn attention to the limitations of preclinical animal trials in determining the safety of drugs in humans."

Mentoring award 2006

Last year we inaugurated the *Nature*/NESTA awards for creative mentoring in science, co-sponsored by Britain's National Endowment for Science, Technology and the Arts. This year we are pleased to announce that *Nature* will be sponsoring awards for high-achieving mentors in two regions: the United Kingdom, again co-sponsored by NESTA, and, later this year, Australasia.

The UK awards are now open for nominations. The closing date is 19 June.

In each region, two prizes will be awarded: one for a lifetime's

achievement in mentoring, and another to an individual in the middle of his or her career. Every nominee has to be nominated by five individuals who between them were mentored over different periods of the mentor's career.

The prizes are intended to celebrate a scientific activity that otherwise tends to be taken for granted. There are many heads of labs whose students have turned into outstanding scientists, but all too often such cases have exemplified survival of the fittest rather than being the product of deliberate nurturing. *Nature* has chosen to favour the latter approach.

Nomination forms and details of the awards can be found at www.nature.com/nature/nestaawards. ■

RESEARCH HIGHLIGHTS

OLFACTION

Map-reading makes sense

Cell **125**, 143–160 (2006)

Understanding how animals perceive smells is hard, particularly because aromas often comprise complex mixes of molecules that vary in concentration and duration. But scientists may make headway with a new map.

Elissa Hallem and John Carlson of Yale University systematically analysed the response of the odorant receptors in the fruit fly *Drosophila*'s antennae to different compounds. This produced the first multidimensional map of an entire olfactory system.

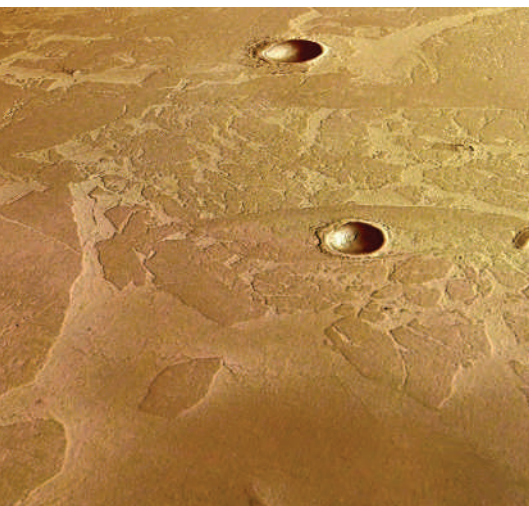
The scientists found the receptors produced both excitatory and inhibitory responses to a diverse range of smells. They also found that receptors with similar odour sensitivity often project to different brain areas.

PLANETARY SCIENCE

Martian mudpack

Icarus **181**, 363–374 (2006)

Last year, high-resolution images of Mars revealed features strikingly similar to pack-ice (pictured). Scientists argued that the pictures showed a frozen lake, and speculated that ice might persist beneath the dust (*Nature* **434**, 352–356; 2005). A new model of how ice sublimates in the arid Martian atmosphere suggests that it would take only tens of thousands of years for the ice content of a thick layer of frozen mud at the site to fall to a few percent. This lingering ice might have cemented in place the pack-ice features, estimated to be about 5 million years old. Konrad Kossacki of the University of Warsaw, Poland, worked with John Murray of the Open University, UK, lead author on the first paper, and others to do the analysis.



Fuel for the future?

Science **312**, 257 (2006)

Researchers in the United States have found a new way to convert small hydrocarbons, which can be obtained from coal or biomass, into the larger molecules needed for transportation fuel.

The process, called alkane metathesis, employs two catalysts to cleave the molecules' ends, and then bond the larger fragments together. For example, hydrocarbons with six carbon atoms yield a mixture enriched in molecules with ten carbons — useful in fuels such as diesel, report Alan Goldman at Rutgers University, New Jersey, Maurice Brookhart at the University of North Carolina at Chapel Hill, and their colleagues.

While the oil holds out, the reaction could also be run in reverse to replace the 'cracking' procedure used in oil refineries (pictured) to break down large hydrocarbons.



AGSTOCKUSA/ALAMY

MICROBIOLOGY

Bug outed

PLoS Pathogens **2**, e28 (2006)

A new disease-causing bug has been found in the lymph nodes of a human patient with a rare immune disorder.

Researchers often struggle to identify bacteria, because most will not grow in the lab. But Steven Holland at the National Institute of Allergy and Infectious Disease in Bethesda, Maryland, and his colleagues managed to culture a bug from a patient with chronic granulomatous disease, where it was underlying a persistent infection.

The bacterium is in the *Acetobacteraceae* family, which was thought to be harmless. Other family members live in plants and soil. The team is investigating whether it might cause more common infections — and have christened it *Granulobacter bethesdensis* after the disease and where it was found.

GENETICS

From 7 to 11

Science **312**, 269–272 (2006)

Geneticists are just beginning to realize that regulatory regions on one chromosome can control the activity of genes on another. Jian Qun Ling of Stanford University and his colleagues report a compelling example.

By tracking down interacting DNA segments, the team found that a control region between the *insulin-like growth factor 2* and *H19* genes on mouse chromosome 7 meets up with a region between the *Wsb1* and *Nf1* genes on chromosome 11. The interaction, which requires a protein called CTCF and controls the activity of the genes on chromosome 11, may help the genes take advantage of transcription-control proteins by bringing them into the same neighbourhood.

MATERIALS

Poke it and see

Nature Mater. doi:10.1038/nmat1632 (2006)

There's no better way to understand the mechanics of a material than to poke it with a sharp tip. So say materials scientists who have used nanoindentation experiments to discover unexpected atomic-scale behaviour.

Graham Cross of Trinity College, Dublin and his colleagues used a technique called field ion microscopy to make a sharp tungsten tip, then observe how pressing the tip into gold deformed a region containing a few thousand atoms. This is roughly the number of atoms that can be modelled in simulations, allowing the first direct comparison between theory and experiment.

ESA/DLR/FU BERLIN

One surprise was seeing the crater heal itself when the tip was withdrawn, a behaviour that simulations do not predict.

CANCER BIOLOGY

Invasion fleet

Cancer Cell **9**, 261–272 (2006)

Groups of tumour cells can invade healthy tissue without cutting ties with their neighbours, say biologists.

Until now, most cancers were thought to become invasive when individual tumour cells reduced their production of a protein called E-cadherin that connects them to other cells and tissue. Like ships casting off their moorings, these single cells take on a form that allows them to migrate and invade other tissues in the body.

Now Gerhard Christofori of the University of Basel and his colleagues have found a new mechanism, where tumour cells producing E-cadherin can also invade. They do so as a group, if cells in the outer edge of the tumour mass also express a protein called podoplanin.

MICROFLUIDICS

Warp speed

Phys. Rev. Lett. **96**, 134502 (2006)

The velocity of fluid in micron-sized channels can be measured by repeatedly trapping and releasing tiny spheres suspended in the flow, physicists in Italy and Britain have shown.

Roberto Di Leonardo of the University of Rome 'La Sapienza' and colleagues at the University of

Glasgow used focused laser beams, or optical tweezers, to trap silica spheres suspended in microfluidic channels. By holding and releasing the spheres, and tracking the intervening motion using digital image processors, the team mapped the velocity field around a larger spinning sphere and a channel outlet. The researchers say the technique could be extended to map velocity fields in three dimensions.

DEVELOPMENT

Assume the position

Dev. Biol. **292**, 317–332 (2006)

Researchers in Europe have watched chromosomes dance around the cell nucleus as mouse embryos go through the earliest stages of growth and development.

Daniele Zink of the Ludwig-Maximilians University of Munich, Pascale Debey of the National Museum of Natural History in Paris, and their colleagues found that chromosomes adopt a characteristic pattern in the nucleus of one-cell embryos (pictured)



after fertilization. The chromosomes underwent two phases of rearrangement as the embryos grew, which may mirror changes in gene activity.

The researchers also showed that transferring the nucleus from an embryonic stem cell into a fresh egg made the chromosomes reorganize into the configuration of a one-cell embryo. This suggests the three-dimensional organization is a necessary part of 'reprogramming' the genome during cloning.

DRUG DELIVERY

Acid put-down

Proc. Natl Acad. Sci. USA **103**, 6460–6465 (2006)

A peptide that spontaneously inserts across plasma membranes can act as a 'nanosyringe' to deliver drugs into cells, researchers led by Donald Engelman at Yale University have shown.

The researchers attached a cargo molecule to the carboxy terminus of the peptide — the end that pokes through to the far side of the membrane. When the peptide inserted itself across the membranes of living cells, the reducing environment inside the cell broke down the linking bond (a disulfide), releasing the cargo into the cell's cytoplasm.

The peptide was derived from the membrane protein bacteriorhodopsin. It was christened pHILIP, for pH (Low) Insertion Peptide, because it inserts only in acidic conditions. Such a property could be useful in treating conditions associated with low pH, such as tumours and stroke damage.

CLOUDS HILL IMAGING/CORBIS

JOURNAL CLUB

Reta Beebe
New Mexico State
University, Las Cruces

Having worked on two generations of spacecraft sent to Saturn, a planetary scientist puzzles over its moon, Enceladus.

When I began working on the Voyager 2 craft, all we knew about the Saturnian moon Enceladus was what could be observed through telescopes. It appeared small and as bright as fresh snow.

Then, in 1981, Voyager 2 passed

near the moon and returned pictures of its surface. As a member of the imaging team, I was surprised by what I saw.

In addition to the craters that we expected, some areas seemed to have been resurfaced by flooding. It wasn't clear how this could happen. Conditions at such a distance from the Sun had to be frigid, and the moon was too small for tidal heating — heat generated by tidal forces stretching and squeezing a body — to produce liquid water.

I liked the theory that an

ammonia–water mixture might act as antifreeze and well up to create a slushy flood. I recall explaining it to a third-grade class. The New Mexico students were already fascinated by the moon whose name looked very much like enchiladas, and the idea of oozing antifreeze drew shivers of glee.

Nearly 25 years later, I am still working for NASA, and have been deeply involved in planning the Cassini–Huygens mission to Saturn. Now its results, published in a March special issue of *Science*

(**311**, 1388–1428; 2006), present me and the children of my 1980s students with new problems to consider.

Cassini–Huygens, having made three flybys of Enceladus, detected ice particles but no ammonia gas, wiping out the antifreeze idea. It has also seen puzzling hot spots that are spewing gas. But how do you make geysers on a cold moon? I hope the Cassini mission will be extended beyond the next scheduled flyby, in 2008, to give us the best chance of finding out.

NEWS

Societies spurn women editors

A rare resignation has focused attention on scientific societies' treatment of women. Theresa Markow, president of the Society for the Study of Evolution (SSE), has stepped down in protest that women were not adequately considered for the editorship of its journal, *Evolution*.

Many think the incident is symptomatic of a wider issue. "I see this as truly problematic," says Patricia Gowaty, an evolutionary biologist at the University of Georgia in Athens. "And it is not unique to the SSE."

The society's rules state that it should create a nominating committee to choose a chief editor. But instead, the society appointed a man after informal queries. It then rejected Markow's request to redo the process. Markow, a geneticist at the University of Arizona in Tucson, resigned the SSE presidency on 18 March, about ten weeks into her term.

"I strongly feel that being inclusive with respect to gender ... is non-negotiable," Markow wrote in her resignation letter. "I cannot serve as president of a society when the council and I share such major contrasts of view."

Officers at the SSE admit the selection process was flawed, adding that they regret Markow's departure. "This resignation has brought into focus the issue of participation of

women," says ecologist Dolph Schluter of the University of British Columbia, Vancouver. "The society should address that." Schluter is the SSE's previous president, and he oversaw the search for an editor.

A committee has been formed to tackle the issue, Schluter points out. However, the society's 2,400 members were not informed of Markow's resignation until 11 April, after *Nature* called for comment.

In the SSE's nearly 60 years, *Evolution* has had only one female editor — Markow, from 1995 to 1999. Other journals are similar. Daphne Fairbairn, an evolutionary biologist at the University of California, Riverside, says she had a "discouraging" experience when she proposed female candidates for an editing position at the *Journal of Evolutionary Biology*, published by the European Society for Evolutionary Biology. "No one else came up with a single woman candidate," says Fairbairn. "When I raised the issue, they looked at me dumbly."

She says society leaders criticized her suggestions unfairly. "They would say 'she is nasty', or 'she didn't do a good job'. No one was going through the men's list and saying those

things." When Fairbairn's term as North American editor ends, the journal will have no female editors.

Juha Merilä, a biologist at the University of Helsinki in Finland and the journal's editor-in-chief, acknowledges there are difficulties appointing women. There are few women at high levels of science in Europe, he says, so the pool of candidates is small. "I am, of course, a little disappointed," he says. "I went

through quite a few names; all declined because of other responsibilities."

The journal of the American Ornithologists' Union, *The Auk*, has not had a woman editor in its 123-year history. But Kimberley Sullivan, an ornithologist at Utah State University in Logan, has a

grant from the US National Science Foundation (NSF) to address such issues, and seems to be making progress. The society's existing fellows pick new fellows at the union's annual meeting, from a slate of nominees. At her first fellows' meeting, Sullivan says women nominees were "trashed". "They started blackballing nominees, with someone saying: 'I was with her on a field trip and she misidentified a bird,'" she says. "It was terrible." The

"They would say 'she is nasty', or 'she didn't do a good job'. No one was going through the men's list and saying those things."

Glint from tenth planet dazzles astronomers

When an object is 16 billion kilometres away but probably only half the width of the United States, how do you work out precisely how big it is? That was the challenge faced by teams estimating the size of UB_{313} , the potential 'tenth planet' in our Solar System, which was first spotted in 2003.

So far two teams have tried, and

come up with different answers.

The first group, led by Frank Bertoldi, an astronomer at the University of Bonn, Germany, measured how much heat UB_{313} emits, and compared it with the body's predicted temperature, based on its distance from the Sun. The team calculated that UB_{313} has a diameter of around 3,000

kilometres (F. Bertoldi *et al. Nature* 439, 563–564; 2006).

Now a team led by Michael Brown, the planetary scientist at the California Institute of Technology whose group discovered UB_{313} , has made its own estimate using the Hubble Space Telescope. The researchers took 28 images of UB_{313} and inferred its size using its

distance from Earth and the width of the object in the image — just 1.5 pixels across. They reckon it is around 2,400 kilometres in diameter, closer to the size of Pluto (see graphic).

That would make UB_{313} unusually bright for its size. Brown's team suggested on 11 April that the body reflects 86% of the light that falls

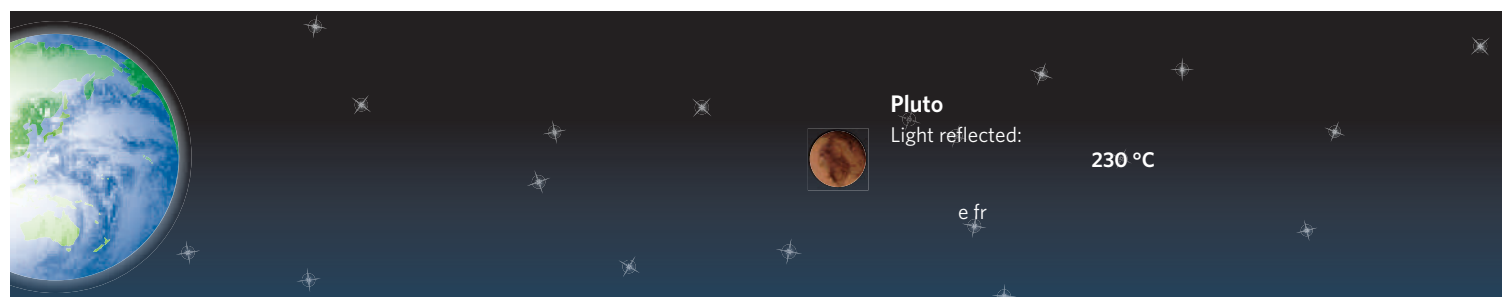




IMAGE100/ALAMY

A man's world? Women researchers accuse some societies of being old boys' clubs.

younger men on the slate came in for the same treatment, she says.

Sullivan presented data from her NSF study at last year's meeting, arguing that the union's tactics were turning it "from an old boys' club into an old men's society". For the first time, the entire slate of prospective fellows was inducted without criticism. "I think the organization has been an old boys' club," says president-elect

Erica Dunn, of the Canadian Wildlife Service in Ottawa. "But the mood is to change."

For Robin Bell, a geophysicist at Columbia University in New York who has an NSF grant to promote women in Earth sciences, change can't come soon enough. "This is a civil rights issue," she says. "We are trying to create women leaders at the institutional level." ■

Rex Dalton

on it. The only object more reflective in the Solar System is Enceladus, a saturnian moon that gets its shiny surface from geysers, which spew water into the object's freezing atmosphere, creating ice on the surface.

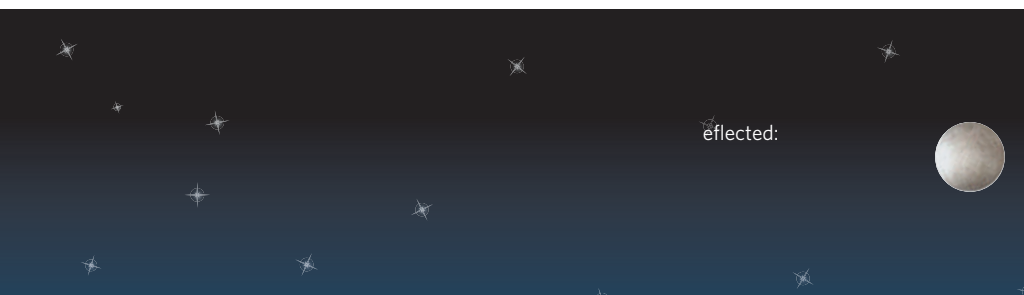
Analysis of light reflected by UB₃₁₃ shows that its high reflectivity is probably due to a coating of frozen methane. But that prompts questions about

how the methane got there, and why cosmic rays, which split hydrogen atoms from methane molecules, have not added a layer of soot to the surface. This process is thought to be responsible for red patches on Pluto's surface, which reflects a maximum of 60% of the light that falls on it.

"Why does it have higher ice coverage than Pluto? We don't

understand it well," says Bertoldi. The body may be leaking methane from its warmer interior. Or it could be because UB₃₁₃ used to be much closer to the Sun, so it may have had a methane atmosphere that froze to its surface as it moved farther out. The same fate probably awaits Pluto as its elliptical orbit takes it away from the Sun. ■

Jim Giles



ON THE RECORD

"ESA is the only space agency to have science operations under way around four planets: Venus, the Moon, Mars and Saturn."

David Southwood, director of the European Space Agency's science programmes, gets a bit confused about what constitutes a planet.

"It didn't come from a toilet on a plane or anything like that."

Charles Glass of the Oakland Fire Department in California discusses a mysterious chunk of ice that fell from the sky and left a metre-sized crater in a city park. Meteorologists say that storms could not have made the ice.

Sources: ESA, Associated Press

SCORECARD



Freedom of speech

Mark Tushingham, a scientist at Environment

Canada, is ordered not to give a public talk about his futuristic novel *Hotter Than Hell*, in which droughts resulting from global warming cause Canada and the United States to go to war over water resources.



H5N1 preparedness

A survey of public-health workers in the United

States reveals that more than 40% say they wouldn't show up at work in the event of an influenza pandemic.

NUMBER CRUNCH

The potential effects on your health of eating chicken nuggets and fries vary depending on the country they're made in — an effect largely dictated by the cooking oil used, according to a recent survey.

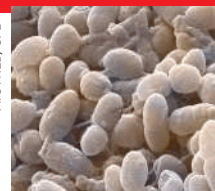
50% of food samples contained levels of *trans* fatty acids linked to an increased risk of heart disease.

38% was the *trans*-fatty-acid content of KFC chicken nuggets bought in the Czech Republic, the highest in the survey.

1% was the *trans*-fatty-acid content of similar nuggets bought in Spain.

Source: Stender, S., Dyerberg, J. & Astrup, A. N. *Engl. J. Med.* 354, 1650-1652 (2006)

A. SYRED/SPL



NOVEL DISEASE-CAUSING BACTERIA
Biologists uncover a threat for patients with a rare immune disorder.
www.nature.com/news

Lakes linked beneath Antarctic ice

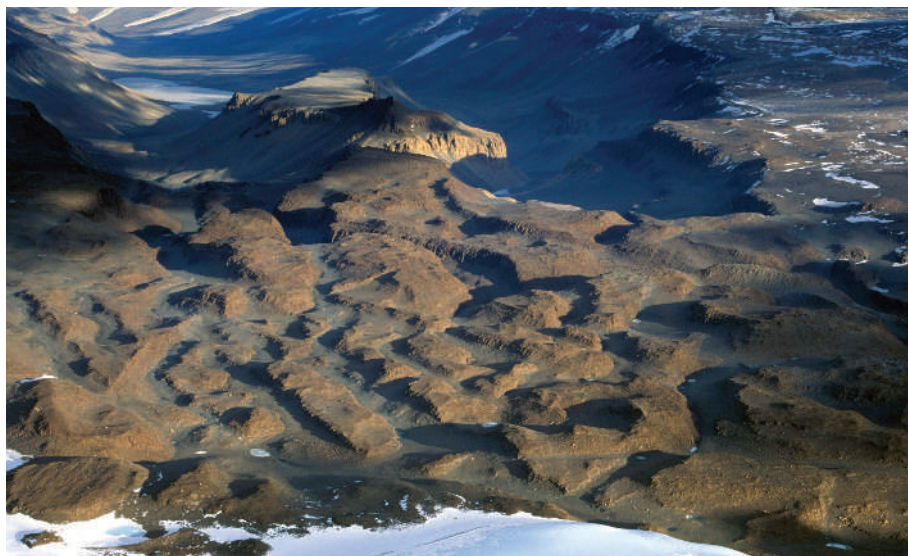
The discovery that lakes hidden deep beneath the Antarctic ice could burst under pressure “like champagne corks” is prompting a rethink of plans to explore the watery world under the continent’s four-kilometre-thick ice sheet. Any flooding might also make its way to the sea, a prospect that has sparked a debate about whether it could affect global climate.

Antarctica’s subglacial lakes have attracted close attention since researchers first realized, around a decade ago, that they may have been isolated for millions of years and so could harbour unique microbial ecosystems. But as awareness has grown, so have the disputes. Many researchers have objected to Russian plans to drill into one of the largest and best-studied bodies of water, Lake Vostok, fearing that it could become contaminated.

The new data, published on page 1033 of this issue, suggest that water from melting ice is pressurizing subglacial lakes, causing rivers with a flow equivalent to that of the Thames to form beneath the ice sheet. “Sooner or later the lakes will go off like champagne corks,” says Duncan Wingham, a physicist at University College London and lead author on the study.

Wingham’s ideas are based on changes in the height of the Antarctic ice sheet revealed by radar devices on the ERS-2 orbiting satellite. A 600-square-kilometre area of East Antarctica dropped by 3 metres in 1997. Around the same time, the ice at two regions 290 km east of this area rose by about a metre (see graphic).

Several lakes are known to exist in these regions, so Wingham and his colleagues con-



M. SIEGERT

Channels in Antarctic rock suggest that subglacial lakes may once have drained into the sea.

clude that the changes are caused by about 1.8 cubic kilometres of water flooding between previously isolated lakes. They suggest that as melting ice flowed into the lakes, changes in pressure forced open tiny channels (see page 1000), causing a flood that lasted for about a year. As the flow subsided, the weight of ice closed the tunnel and sealed the lake again.

About 150 subglacial lakes have already been identified in Antarctica, but thousands more may exist, and the same process could affect them all. Wingham doubts whether all the lakes are connected, but says that some are linked like “beads on a string” by channels that periodically burst open.

If lakes such as Vostok were not isolated, they would be less interesting to researchers, although even if they are linked they could still share a unique ecosystem. And if they are connected, the risks associated with drilling increase. “If we accidentally contaminate one lake, will we contaminate a series of lakes via hydrological conduits?” asks John Prisco, a microbial ecologist at Montana State University in Bozeman, who studies subglacial lakes. He says Wingham’s results should be considered when planning exploration projects.

US and European researchers are debating how best to minimize the risk of contamination, and don’t expect to be ready to drill into a lake before the end of the decade. But a Russian team plans to enter Lake Vostok as early as

2007. That worries others, because the borehole, which currently ends 100 metres above the surface of the lake, contains drilling fluid that could potentially contaminate the water.

Valery Lukin of the Arctic and Antarctic Research Institute in St Petersburg, head of the Russian drilling efforts, says the Wingham

“Sooner or later the lakes will go off like champagne corks.”

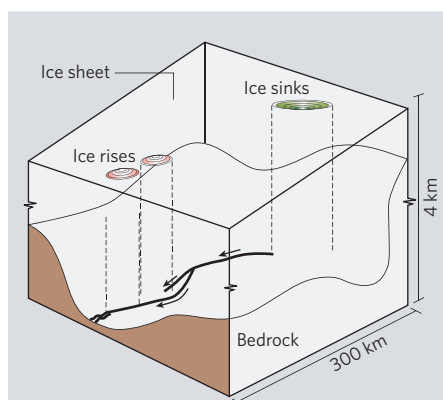
paper “will not influence plans to drill into Vostok”. Lukin says his team is confident that the pressure in Lake Vostok will

force water up into the borehole, where it will freeze and act as a cap to prevent contamination.

Researchers are also asking what would happen if a large subglacial lake discharged into the sea. Antarctica’s edge has twisting channels in the rock that researchers believe were formed in the Miocene, 13 million years ago, by the drainage of subglacial lakes. David Marchant, a geologist at Boston University in Massachusetts, says the water would have drained into the ocean, where it could have disrupted circulation patterns and influenced global climate. So this process could conceivably occur again.

“It depends on the total volume of water that drains into the ocean and how fast it gets there,” he says. To determine the risk, Marchant says that a better understanding of the timing and size of the Miocene floods and the climate changes of the period is needed. Researchers will also need to work out which modern lakes are connected.

Jim Giles



Satellite data showed ice sinking above one subglacial lake, while rising at two spots 290 km away.

The researchers conclude that water must have flowed from the first lake into two others downstream.

Wanted: cancer boss

The White House is looking for a new head for the National Cancer Institute. What kind of person does it take to run a \$4.8-billion research powerhouse? *Nature* asked some top cancer researchers for their thoughts.

Donald McDonald

University of California, San Francisco

It is much easier to describe the perfect person to be director of the National Cancer Institute (NCI) than to find him or her. That person would need to balance the needs, hopes and challenges of state-of-the-art clinical oncologists with those of basic scientists. The ideal candidate must understand the inner workings of the NCI as the pace-setting institute that it is, but also as one of 27 separate institutes and centres — with competing identities, missions and needs — that make up the National Institutes of Health (NIH). Also, this person will have to address the challenges faced by cancer research, like all biomedical research, with declining funding and the changing political climate.

One approach would be to look for someone within the NIH — someone who recognizes all of these problems yet is not daunted by the financial and political conditions. An example would be the current NCI deputy director, John Niederhuber.

Bert Vogelstein

Johns Hopkins University School of Medicine, Baltimore, Maryland

I would suggest Bill Gates as the new director, for two reasons. First, he is one of the few people who could make up shortfalls in funding. Second, and more seriously, he would bring a business perspective that is sorely needed. The NCI is riddled with old programmes that are viewed almost as entitlements and have hindered past

directors from realizing their visions. These vestiges of the past drastically limit the agency's ability to support novel initiatives and, most importantly, new investigators. Just as a business evaluates each of its products in an effort to maximize profits, each NCI programme should be subjected to a "zero-based budgeting" analysis based on scientific or clinical productivity and the likely impact on disease.

Mary-Claire King

University of Washington, Seattle

In my view, the next NCI director should have direct experience of NCI-funded, extramural research in his or her own lab; experience of translating research into patient care; and knowledge of how the NIH and NCI work. I would begin my list with the current directors of the 39 NCI-designated Comprehensive Cancer Centers. These people are respected, and have worked for cancer patients and researchers for their entire careers.

Waun Ki Hong

University of Texas M. D. Anderson Cancer Center, Houston

The NCI director must have a strong track record as a cancer researcher. This person needs a really strong vision and passion for cancer research. She or he must be an excellent advocate with keen political skills. And they must be highly respected and recognized by their peers in the cancer community. With these qualities in mind, I think any of the following people would make a fine director: Phillip Sharp of the Massachusetts Institute of Technology; Max Wicha of the University of Michigan; Harold Moses at the Vanderbilt-Ingram Cancer Center; and Martin Abelson at Johns Hopkins University.

Hellmut Augustin

University of Freiburg, Germany

Commenting on an American issue from a European perspective is like interfering in domestic affairs — usually not well appreciated. On second thoughts, it is quite tempting. It should be a strong personality who is well rooted in basic as well as translational oncology research, a scientist with a vision beyond

THE NCI AT A GLANCE

Location: Bethesda, Maryland, with a satellite facility in nearby Frederick.

Budget: \$4.8 billion annually, making it the largest institute in the National Institutes of Health.

Purpose: Among many others, to conduct leading research into the causes and treatment of cancer. Andrew von Eschenbach, NCI's outgoing head, has set a goal to eliminate suffering and death from cancer by 2015. Many researchers see this as unrealistic.



"Thirty years of cutting-edge research have given Harold Varmus a deep understanding of the biology of cancer."

NEWS.COM

the tumour-cell-centric view, and probably (but not necessarily) an MD or an MD/PhD. The job needs someone from outside who can motivate and develop a new corporate identity at the agency, a person of the highest degree of scientific integrity, and an efficient administrator with excellent communication skills. The first person who comes to mind who would fulfil such criteria is Carl-Henrik Heldin at the Ludwig Institute for Cancer Research in Uppsala, Sweden. But why not a female scientist? Margaret Kripke of M. D. Anderson is a member of the President's Cancer Panel. How about a strong basic scientist with an industry background, such as Mark Fishman of the Novartis Institutes for



C. SPIELMANN

NEWS.COM



"Bill Gates would bring a business perspective that is sorely needed."

BioMedical Research? Or a bona fide cancer researcher: Bert Vogelstein, Doug Hanahan, Robert Weinberg of the Whitehead Institute or...? Well, a European should probably not interfere in domestic affairs!

Douglas Hanahan University of California, San Francisco

I would choose Harold Varmus, currently president of the Memorial Sloan-Kettering Cancer Center. Thirty years of cutting-edge research have given Varmus a deep understanding of the biology and genetics of cancer. Formerly director of the NIH, he knows about leading a major government agency. Now, as president of Sloan-Kettering, he has assembled

an impressive team in basic, translational and clinical cancer research and cancer care, and has recruited world-class scientists and physicians to his institution. This combination of research excellence, grounding in medicine and experience in relevant institutional leadership has uniquely prepared Varmus to guide the NCI and the cancer research community.

Marc Lippman University of Michigan School of Medicine, Ann Arbor

I think the ideal candidate should be seen by scientists and others in the field as scholarly, original in their thinking, and able to capture the enthusiasm and involvement of the many diverse constituencies that make up the community of people who dedicate their lives to eradicating cancer. Here is an eclectic set of names who I think meet all of those criteria: Mina Bissell of the Lawrence Berkeley National Laboratory; Isaiah Fidler of the University of Texas M. D. Anderson Cancer Center; Larry Norton of Memorial Sloan-Kettering Cancer Center; and Philip Pizzo of Stanford University.

Interviews by Jacqueline Ruttimann and Meredith Wadman

US cancer funding creates business as well as science

One in four of the scientists funded by the US National Cancer Institute (NCI) who seek patents have started their own companies, says a study released last week.

This is a higher rate of entrepreneurship than previously measured, says David Audretsch, an economist at Indiana University and a co-author of the report. The firms' success has yet to be analysed.

In a project funded by the Kauffman Foundation of Kansas City, Missouri, Audretsch and colleagues studied 1,700 scientists funded by the NCI from 1998 to 2004. Nearly 400 had sought US patents, and 25% of the 140 who responded to the survey had started a company.

The study reinforces the critical role that government funding plays in innovation, says Carl Schramm, president of the foundation.

UK unleashes watchdog to enforce research integrity

Britain has launched a panel to oversee integrity in biomedical research, with the aim of encouraging what they term a

'zero-tolerance' approach to misconduct.

The UK Panel for Research Integrity in Health and Biomedical Sciences, which launched on 12 April, will advise and train those investigating misconduct allegations, and support whistleblowers in universities and the National Health Service. The project will have a budget of £200,000 (US\$350,000) and initially run for three years.

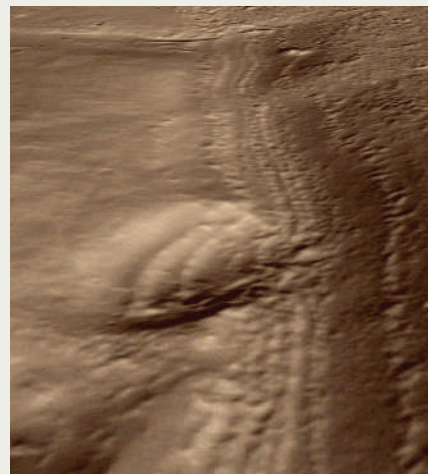
Plans for the panel were criticized when

revealed last year (see *Nature* 434, 263; 2005), after organizers said they would accept funding from the pharmaceutical industry. Universities UK, the umbrella organization for British higher-education institutes, says that the Association of the British Pharmaceutical Industry (ABPI) is helping to fund the project, but that the activities of the panel will be independent of the ABPI and all other funders.

Latest eye on Mars rises to the occasion

Martian photography just gets better all the time. This close-up of a ridge in the planet's southern highlands is one of the first images returned by the powerful High Resolution Imaging Science Experiment (HiRISE) camera on the newly arrived Mars Reconnaissance Orbiter. The vertical relief in this perspective view is roughly what the eye would see.

Running up from the bottom of the frame, the ridge has deformed the edge of a small impact crater. The first HiRISE test images were taken on 24 March from an altitude of about 2,500 kilometres, and show objects as small as 7.5 metres across. Once the probe has settled into its final orbit in November, the camera will be about ten times closer to the surface, and will routinely resolve objects well under a metre in size.



NASA/JPL/UNIV. ARIZONA/USGS

H. FERNANDO, ARIZONA STATE UNIV.



Hands off the coral, says fisheries ministry.

Sri Lankan signs warn coral thieves of nature's wrath

They won't bring back the coral that's been taken, but new billboards in Sri Lanka might help stop further illegal collecting.

Visiting the coast at Akurella after the devastating tsunami of 26 December 2004, Harindra Fernando, a Sri Lankan fluid dynamicist at Arizona State University in Tempe, found that spots where coral had been removed sustained greater damage (see *Nature* 436, 1071; 2005).

Now, this sign has been erected along the same coast. It reads "Let's refrain from mining corals that control beach erosion," in Fernando's translation, and is signed by the Coast Conservation Department of the

Ministry of Fisheries and Aquatic Resources. It sits near a place where the tsunami shot through a gap in the reef and washed a train off its tracks. Some protective rock walls have also been built in the area, Fernando says.

Satellite monitor of the seas confirmed for 2008

International partners are gearing up to make more satellite measurements of sea levels.

On 10 April, European and US agencies signed an agreement to launch the Ocean Surface Topography Mission in 2008. Known as Jason-2, it will follow the Jason satellite, launched in 2001, and the Topex/Poseidon mission, launched in 1992. Both have provided key measurements on sea-surface height, in part to monitor the effects of climate change.

France's National Centre for Space Studies is providing the spacecraft and several science instruments. NASA is contributing instruments and the launch. The National Oceanic and Atmospheric Administration and the European Organisation for the Exploitation of Meteorological Satellites will run the mission once it is in orbit.

A good spin helps crack puzzle of jumping eggs

Hard-boiled eggs can jump, a team of Japanese researchers has shown. All it takes, say Yutaka Shimomura and colleagues at Keio University, is a good spin.

A spinning egg will spontaneously rise up from lying on its side to standing on its end. Shimomura, along with physicists at the University of Cambridge, had previously worked out why — friction grips the egg at the pivot point and turns some of the spin into uplift. The team predicted that the forces involved could also make an egg leap a tiny bit into the air.

But does it really happen? To check, the team used an ovoid-shaped metal egg, to avoid making too much of a mess. They placed their egg horizontally on a slab of polished copper, started it spinning and filmed it. Three different methods of detection all showed that the egg made tiny jumps before rising to a fully upright position.

What works for the metal egg also works for a real hard-boiled egg, the team reports in *Proceedings of the Royal Society A* (T. Mitsui *et al.* doi:10.1098/rspa.2006.1718; 2006).

SPECIAL REPORT



Counting the dead

Twenty years after the worst nuclear accident in history, arguments over the death toll of Chernobyl are as politically charged as ever, reports **Mark Peplow**.

No more than 4,000 people are likely to die as a result of Chernobyl. That was the conclusion released by the United Nations and the governments of Ukraine, Belarus and Russia in September last year, in the most comprehensive assessment of the accident so far.

But despite promising “definitive” answers the report, based on two decades of research, has done little to resolve the debate over Chernobyl’s impact. The estimate drew howls of protest from environmental groups, which accused the UN’s Chernobyl Forum of a whitewash. And scientists whose work is cited in the report are concerned about how their figures were presented, pointing out that the true cost of the disaster will not be known for decades to come, if ever.

Chernobyl’s runaway nuclear reaction in 1986 was triggered by a faulty safety test, and exacerbated by a design flaw that caused a catastrophic temperature rise in the reactor core. This caused about 6.7 tonnes of radioactive material to spread for hundreds of kilometres around the site. Two of the most significant elements in the chemical cocktail were iodine and caesium, and the most devastating effects were seen within a few tens of kilometres of the reactor, in the region where Ukraine, Belarus and Russia meet.

In the confusion that followed, local families continued to graze and milk their cattle on contaminated land. Many predicted that the accident would cause hundreds of thousands of cancers.

Cooking the books?

The forum reports that the fallout has since caused about 4,000 cases of thyroid cancer, mainly in children and adolescents. But just 15 of these patients have died. Along with highly exposed rescue workers who brought the reactor inferno under control, 62 deaths have been attributed directly to the accident so far.

The report also says that there has been no significant increase so far in the incidence of other cancers. In total, it said, “up to 4,000 people” may ultimately die as a direct result of the disaster — much lower than previous estimates.

That conclusion upsets many. “The report gave a completely misleading view of the health consequences of the accident,” says Ed Lyman, a nuclear-power expert with the Union of Concerned Scientists in Washington DC.

The argument stems from uncertainty about the health effects of low doses of radiation. Evidence from survivors of the atomic

bombs dropped on Nagasaki and Hiroshima has allowed researchers to predict the effects of high doses. But as the exposure drops, so does the understanding of its impact. Last year, a working group of the US National Research Council concluded that even tiny amounts of radiation cannot be considered safe.

In contrast, the forum based its headline figure on just the 600,000 people exposed to the most radiation, predicting that roughly 4,000

of them will die as a result. The full report acknowledges that, of 6.8 million others living further from the explosion who received a much lower dose, Chernobyl will kill another 5,000 — more than doubling the projected death toll.

But this is not mentioned in the report’s 50-page summary or the accompanying press release.

The figures come from a study published in 1996 by Elizabeth Cardis of the International Agency for Research on Cancer in Lyon, France. “I was very shocked that they were quoting figures we had found ten years ago,” says Cardis. “I didn’t expect the numbers to be picked up and used in a press release without qualification.”

Even the forum’s chair, Burton Bennett, former head of the Radiation Effects Research

“I was sick of seeing wild figures being reported by reputable organizations and attributed to the UN.”

**IRAN GOES NUCLEAR**

news@nature.com
puts Iran's efforts to
enrich uranium into
perspective.

www.nature.com/news

REUTERS/IRNA

I. KOSTIN/CORBIS

Foundation in Hiroshima, Japan, says caveats about the figures should have been clearer. But he points out that the extra 6.8 million people were, on average, exposed to a radiation dose of just 7 millisieverts — little more than the natural background delivers in a year in most parts of the world.

Melissa Fleming, a press officer working at the International Atomic Energy Agency in Vienna, who helped coordinate the report's publicity, says the scientists involved checked the press material. But she admits a decision was made to focus on the lower 4,000 figure, partly as a reaction to the inflated estimates of past decades. "I was sick of seeing wild figures being reported by reputable organizations that were attributed to the UN," she says. "It was a bold action to put out a new figure that was much less than conventional wisdom." The figure has been removed from the final summary, however, published this month.

Uncertain legacy

Some of the higher death tolls come from those who believe that estimates of those potentially affected should extend to everyone in Europe (see map). In a report commissioned by Green Party members of the European Parliament, radiation scientist Ian Fairlie calculated that of the hundreds of millions of people who could

possibly have received any radiation at all from Chernobyl, 30,000–60,000 could die as a result (see <http://tinyurl.com/rpzsq>).

Cardis is also about to publish a study of the pan-European impact. She concludes that, of 570 million people in Europe at the time, 16,000 will ultimately die as a result of the accident — 0.01% of all cancer deaths. But she says it will be virtually impossible to assess the ultimate death toll. Cancer causes about a quarter of all deaths in Europe, so weeding out those cases triggered by Chernobyl cannot be done with statistical confidence. "We'll never be able to say whether we were right or not," she says.

Some researchers also take issue with the report's conclusion that there are no hereditary effects in children born after the disaster. "The fact that we haven't seen anything doesn't mean there isn't an effect," says Cardis. "It's just too early to see an increase."

Until Chernobyl survivors' children and grandchildren grow up, the only way to assess such effects is to look for DNA mutations. Yuri Dubrova, a geneticist at the University of Leicester, UK, has found increased genetic changes in the children of irradiated parents — but the fingerprinting technique he used only allowed him to look at non-coding regions, known as junk DNA (Y. E. Dubrova *et al. Nature* **380**, 683–686; 1996).

"The fact that we haven't seen anything doesn't mean that there isn't an effect. It's just too early."

He argues that coding regions are also likely to be affected. DNA chips currently being developed will allow mass screening of samples from thousands of people, he says, and could pick up effects too small to have been spotted so far. "The real necessity now is to organize proper blood-sample banks," he says.

Many are calling for research to track effects such as cancer or infertility in subsequent generations (see page 993). But although such monitoring is seen as essential, some worry that wrangling over death tolls and radiation risk is hampering survivors' recovery.

Against apathy

"What we'd like people to take away is not the numbers game," says Louisa Vinton, who manages Chernobyl projects at the United Nations Development Programme (UNDP). Rough estimates of death tolls allow governments to set up policies to manage the future health effects. But she says the focus on the figure of 4,000 obscured a more important message of the report: that myths about the threat of radiation have created a "paralysing fatalism" among residents of affected areas.

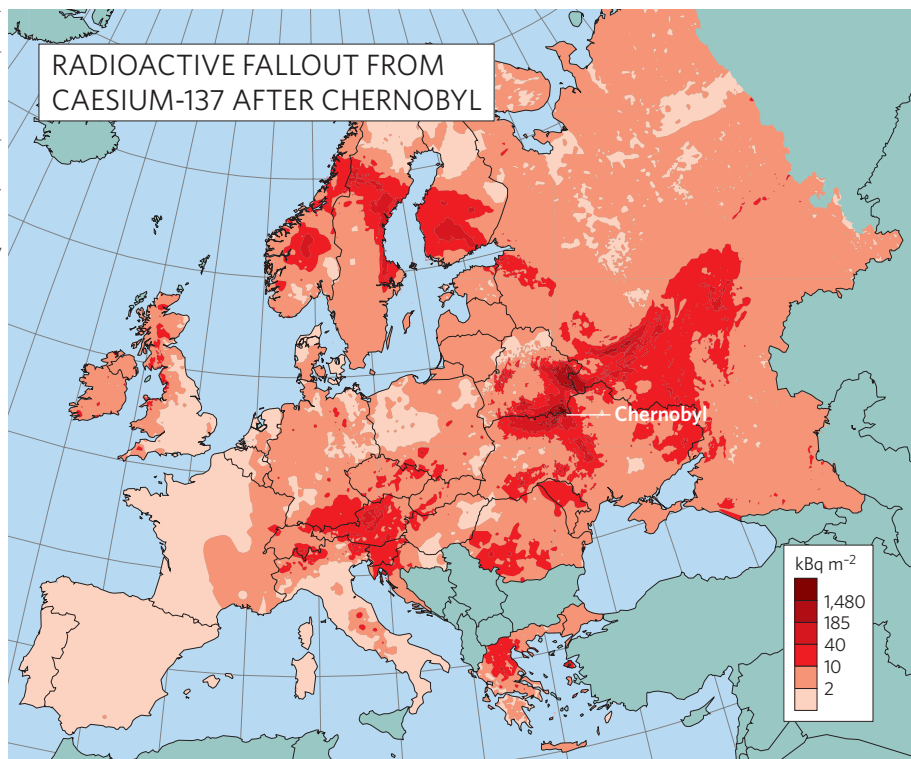
Instead Vinton hopes that Chernobyl's legacy can be seen as a social problem. The UNDP says that Chernobyl's most serious impact was on the mental health of about 7 million people labelled as victims of the accident. Aid from the governments of Russia, Belarus and Ukraine has created a culture of dependency, it argues, which may have encouraged exaggerated fears of ill-health.

"We go to communities where people have just given up," says Vinton. Poor diet, lack of exercise and smoking are all linked to such apathy. She admits it is hard to separate the stress of Chernobyl from the effects of collapse of the Soviet Union. But Chernobyl aid has drained resources. In 1991, Belarus spent 22.3% of its budget on aid; today that figure is still 6%.

The forum's report recommends that governments reassign the cash towards regenerating the region and developing infrastructure.

The hottest spots will be radioactive for centuries. But the most prevalent isotope, caesium-137, has a half-life of about 30 years and scientists estimate that much of the abandoned area will become habitable over coming decades. The 30-kilometre exclusion zone around Chernobyl is likely to remain off limits. But the report suggests that in other evacuated areas roads should be rebuilt, and people encouraged to start farms and businesses.

Projects to build hospitals and schools are already helping people to help themselves, says Vinton: "I've seen the before and after. People stand up straight and bubble over with excitement about the next plan."



J. SMITH & N. A. BERESFORD CHERNOBYL: CATASTROPHE AND CONSEQUENCES (PRAXIS, CHICHESTER, 2005)

WHEN THE PRICE IS RIGHT

Once touted as too cheap to meter, nuclear power has become too costly to build. But the economics may be shifting, finds **Jim Giles**.

On the east coast of Britain sits one of the most convincing arguments for ending the age of nuclear power. The Sizewell B reactor is not dangerous: since it opened in 1995 it has never been the subject of a serious security scare. Nor is it unreliable. Quite the reverse: at almost 1,200 megawatts, it is Britain's most powerful single nuclear reactor and is responsible for supplying 3% of the country's electricity needs. But Sizewell B is expensive. So expensive that no private investor would ever touch such a project.

For nuclear experts, the story of Sizewell B is a familiar one. After the longest public inquiry into a construction project that Britain had ever seen, work began in 1987. It took eight years to come online. The budget was revised upwards three times over that period, eventually coming in at more than a third over the £2 billion (US\$3.3 billion) quoted in 1987. When the British government reviewed the project in 2002, it estimated that, when the costs of financing, building, running and decommissioning Sizewell B were fully accounted for, the average cost of every kilowatt hour (kWh) of electricity produced over the plant's 40-year life would be six pence — two to three times more expensive than power generated by modern gas-fired stations.

Running down

Nuclear power stations account for a fifth of the electricity generated in the United States and a third of that generated in Europe. But they are getting old. Since the 1979 accident at the Three Mile Island plant in Pennsylvania, orders for new reactors in the United States and Europe have reduced to a trickle. Decisions on how to replace the existing plants need to be made within the next ten years. And although renewable sources such as wind are looking increasingly attractive, large central power stations can only realistically be powered by nuclear fission, coal or gas.

Almost all recent studies of nuclear energy have found that gas- or coal-fired replacements would be much cheaper. Given this underlying lack of competitiveness, why bother taking on board the associated risks of terrorism and weapons proliferation that come with the technology, not to mention the

displeasure of many citizens? The answer is that when the downsides of fossil fuels — including, but not limited to, their carbon dioxide production — are totted up, nuclear power begins to look more attractive. Some economists are even starting to place bets on a nuclear renaissance.

The current economic picture is persuasively summed up in the 2003 nuclear-economics report from the Massachusetts Institute of Technology (MIT)¹. After considering the cost of building the plant, buying fuel and operating the reactor, and finally disposing of the waste and decommissioning the facility, the MIT team placed the cost of nuclear electricity at 6.7¢ per kWh. Gas came in at 3.8–5.6¢ per kWh, depending on wholesale gas price, with coal somewhere in the middle of that range. A 2005 report by the UK Royal Academy of Engineering² put nuclear costs on a par with coal and gas, but used some unreasonably favourable economic assumptions.

Much of nuclear power's expense comes from construction costs, and the debts that must be incurred to pay them (see chart). A 1,000-megawatt gas-fired plant could, in favourable circumstances, be built in a year or so for \$400 million, but a 1,000-megawatt nuclear reactor is likely to take five years to build and to cost between \$1.5 billion and \$2 billion, depending in part on where it is sited. The long construction time drives the price up by increasing the amount of interest that must be paid on the money borrowed for the project.

And the length of the gestation isn't even

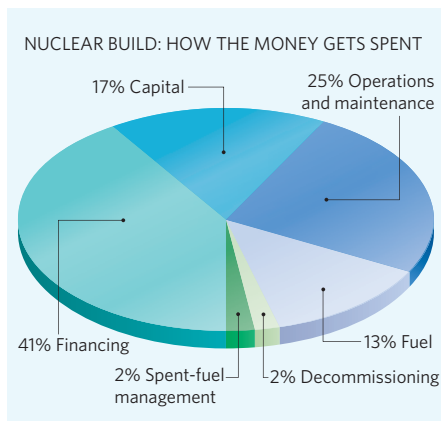
predictable; it depends in part on "how slick the lawyers are", notes Donald Jones, an energy economist at RCF Economic and Financial Consulting in Chicago, Illinois. If opponents of nuclear power raise legal challenges, costs mount up quickly. "Halting construction for two years in the middle adds 15% to the final cost of electricity," Jones says. In 2005, hoping to encourage the industry by offsetting this risk, the US Energy Policy Act offered energy companies \$500 million of coverage for losses due to construction delays. Yet US investors remain wary of nuclear projects.

Balancing act

Once a nuclear plant is running, operational costs are relatively low. But there are two important exceptions: storage of radioactive waste and, at the end of the reactor's life, decommissioning costs. In Britain, estimates of the funds required to clean up the country's 20 civil nuclear sites have frequently been revised upwards. Just last month, for example, officials increased the total predicted bill from £56 billion to £70 billion. Some schemes to deal with waste costs are in place — the United States has a levy on nuclear electricity that will fund the country's planned waste-storage facility at Yucca Mountain in Nevada (see page 987). But uncertainty about these costs continues to worry investors.

The economic picture hasn't stopped all construction of nuclear power plants. India, China and Russia are building a handful of reactors, for example, but all are government-funded projects. It is in Europe and the United States, with their largely deregulated energy markets, that the economic arguments have bitten deepest. Finland is the only Western nation to have a new nuclear reactor under construction, but its Olkiluoto station, due for completion in 2009, makes economic sense only because of an unusual funding mechanism: local industries are paying for the plant in return for a contract from the operator that guarantees them low-cost electricity. This is not a model that can easily be exported.

France may also decide to order a new reactor, possibly this year. But again the situation is different from that in most countries. France's experience with the technology means that investors will loan money for projects at a





Core product: Britain has built no nuclear power stations since Sizewell B, in part because of the expense.

cheaper rate than elsewhere. A French 2003 study, for example, put the price of future nuclear electricity at 3.5¢ per kWh — the lowest figure in any of the recent reviews, with the exception of estimates for Olkiluoto.

When cost comparisons are extended beyond current prices and business practice, however, nuclear looks like a feasible option even beyond the borders of France. For a start, although the nuclear industry faces some unique challenges, the finances of its two major competitors are also looking a little troubled. Take gas: wholesale prices have increased fourfold over the past six years. That pushes the price of electricity from gas-fired stations to around 15% below nuclear.

That gap will shrink further if industry forecasts about the efficiency of modern plants are accurate. Nuclear lobby groups claim new plants could probably be built in four years, not five. Independent experts are aware that the industry has a record of what is euphemistically known as 'appraisal optimism'. But the MIT study, which worked with costs larger than those the industry usually uses, acknowledges that improvements are "plausible". If this were so, the cost of nuclear electricity would come down to 4.2¢ per kWh, making it competitive with gas and coal.

Fair exchange

Factor in the cost of greenhouse-gas emissions, and things look even better for nuclear. In Europe, where emissions from industry are already regulated and traded on a carbon market, prices are currently around \$30 per tonne of carbon dioxide. At that price, says Paul Joskow, an economist and participant in the MIT study, nuclear is competitive with coal and gas, at least if the price of the latter remains high. This remains the case even when the cost of the carbon burned while the plants are built and their fuel mined and processed is taken into account.

This broader analysis, which the nuclear industry is understandably keen to promote, boosts its standing. But where should the process of extending the cost comparisons end? Environmental groups say that going as far as carbon prices and no further means taking into account all the costs of other generators while leaving out costs specific to nuclear, such as lowering the barriers to nuclear proliferation. Assessing that argument takes the calculations on to less certain ground, although a smattering of studies have attempted to quantify some of the issues. Inasmuch as anything can be said for sure, however, it seems that the more inclusive approach may improve the case for nuclear.

Take disaster liability. In Britain, the amount that reactor owners have to pay out in the

G. SMITH/CORBIS

event of an accident is limited to £140 million (\$250 million). US industry contributes to a pool of money that ensures that up to \$10 billion is available. Neither figure would be anything like sufficient should a disaster on the scale of Chernobyl occur (see page 982). The extra cost would have to be picked up by the taxpayer, so “in essence this is government-subsidized insurance”, says Matthew Bunn, a nuclear expert at Harvard University. If nuclear were forced to insure itself on the open market, it would find it impossible.

Government crutch?

But how big a subsidy this actually is remains unclear, because the real risk of catastrophic accidents is unknown. Estimates can be generated by looking at the frequency of previous accidents and how the associated costs compare with events for which the insurance industry is prepared to provide cover. The MIT study suggests that the subsidy amounts to just \$3 million per plant per year — a tiny figure when reactors produce \$500 million of electricity annually. “In terms of the impact on the cost of electricity it’s lost in the noise,” says Richard Lester, an author on the MIT study.

What’s more, nuclear is not the only industry that benefits from subsidized insurance. A major explosion at a depot handling liquid natural gas could produce a bill well beyond the scope of the owner’s cover. So would the wall of water let loose from a hydroelectric dam destroyed by an earthquake. “There is an implicit assumption that the government would step in,” says Lester. “Everything has an insurance limit.”

Some other costs are simply unquantifiable. The European Union’s Externe study³, which has been running since 1991, provides perhaps the fullest accounting of what economists call ‘externalities’ — costs that the people directly involved don’t end up paying. Externe’s audit assigns nuclear extra environmental costs of 0.2–0.8¢ per kWh, mostly derived from air pollution attendant on the plants’ construction, mining and transport of fuel, and decommissioning. The figures for fossil fuels, which include damage to the climate as well as air quality, are much higher — up to 18¢ per kWh for coal. Yet even when Externe’s comprehensive analysis is considered, some things are still unaccounted for. “We can’t include terrorism issues,” says Anil Markandya, an economist at

“We can’t include the cost of terrorism issues. We don’t have a handle on how to quantify that.”

— Anil Markandya



Buried costs: uranium is a small part of a power plant’s expenses, but mining it generates carbon dioxide.

the University of Bath who works on Externe. “We don’t have a handle on how to quantify that.” There is also the cost that would be incurred were an unstable nation to develop nuclear weapons by buying nuclear-reactor technology. “The contribution of the civil nuclear system to proliferation is impossible to monetize,” says Bunn. “But that would be the biggest externality.”

Unstable fuel

Such costs, even if they cannot be quantified, do not apply only to nuclear. “If you’re concerned about nuclear safeguard costs you have to look at the costs of other sources,” says William Nuttall, a nuclear expert at the University of Cambridge, UK. Putting a figure on the Western military spending associated with maintaining fossil-fuel supplies from the Middle East is a politically contentious task. But various estimates, from tens of billions of dollars a year to more than a hundred billion, suggest there is a hidden subsidy for oil prices that might top 10%.

These arguments, although vital for policy-makers wondering what to encourage, will not on their own influence investors’ decisions. But a final point in favour of nuclear comes from a source that the money men are used to listening to. Portfolio theory is an established way of generating a mix of investments that creates maximum return for a given level of risk. This, says Shimon Awerbuch, an economist at the University of Sussex, UK, is exactly how governments should approach energy decisions. “Talking about generating cost without also talking about financial risk is like watching a movie with the sound turned off,” he says. “You miss a big part of the story.”

In the case of electricity generation, ‘risk’ concerns the chance that fuel prices, be they uranium or gas, will go up. Hikes in oil prices have a similar knock-on effect to those of energy prices more generally — they reduce gross domestic product. The real cost of a fuel source, says Awerbuch, needs to take such risks into account. That is bad news for

sources whose price fluctuates, such as gas, and good news for nuclear, as uranium costs are reasonably steady, and likely to remain so unless there is an unparalleled boom in plant building.

Awerbuch’s approach is to analyse the current mix of fuel sources in the economy to see what levels of risk governments are implicitly willing to accept. He then searches for other mixes that deliver the same risk at less cost. When trying out new combinations, something surprising can happen: adding an expensive non-fossil-fuel source such as nuclear or wind can actually decrease the overall cost. Nuclear lowers exposure to price hikes, and that lets planners simultaneously invest in riskier but cheaper sources such as gas. That additional gas more than compensates for the more expensive nuclear power, so overall prices fall. Although most of Awerbuch’s work focuses on wind⁴, his analysis also suggests that the steady price of uranium means nuclear should be retained as part of a healthy mix of generation sources.

When Awerbuch’s way of looking at energy is combined with the recent rises in gas prices and, more significantly, the new carbon markets in Europe and the United States, another round of nuclear build seems a realistic possibility. A straw poll of nuclear experts shows they are starting to be convinced. Bunn used to offer straight bets against new nuclear construction starting in the coming decade. But put all these changes together, he says, and he might need to start offering odds. “Over 15 years,” he adds, “I might switch my money to the other side.” ■

Jim Giles is a senior reporter at Nature. See Editorial on page 969.

1. *The Future of Nuclear Power: An Interdisciplinary MIT Study* (Mass. Inst. Technol., 2003); published online <http://web.mit.edu/nuclearpower>
2. *The Costs of Generating Electricity* (R. Acad. Eng., London, 2004); published online http://213.130.42.236/wna_pdfs/rae-report.pdf
3. *Externe: Externalities of Energy* (European Commission, 1995).
4. Awerbuch, S. *Mitigation Adapt. Strateg. Glob. Change* (in the press).

STOCK CONNECTION DISTRIBUTION/ALAMY



Experiments in a tunnel at the Yucca Mountain nuclear-test site in Nevada aim to establish whether nuclear waste could safely be dumped here.

DOE

FORWARD PLANNING

The global future of nuclear power may rest in large part on local politics, reports **Geoff Brumfiel**.

Like 83% of the state of Nevada, the land to your right as you head up from Las Vegas towards Reno is owned by the federal government. A military airstrip for unmanned aerial vehicles and a nuclear-weapons test range stretch into the distance; enormous concrete tubes that once held MX missiles lie baking in the sun. The only signs of civilian life are a state prison, a brothel and a few small clusters of air-conditioned trailer homes.

It is here in Nye County — where 38,000 souls occupy 47,000 square kilometres — that the US Department of Energy would like to bury more than 70,000 tonnes of highly reactive nuclear waste. The federal government says that Yucca Mountain, a low peak on the western edge of its nuclear test site, some 100 kilometres northwest of Las Vegas, is one of the safest places in America for spent fuel to make its several-hundred-thousand-year journey to harmlessness.

Local residents want jobs, so they are not wholly opposed to the repository. But other citizenry of Nevada, many of whom remember the consequences of above-ground nuclear-weapons testing in the 1950s and 60s, are deeply sceptical. And so are their elected representatives. “The state of Nevada is committed to stopping this thing by any legal means possible,” says Steve Frishman, a geologist and the technical

policy coordinator for the state government’s Agency for Nuclear Projects.

Virtually every democratic nation that has embarked on a programme for the disposal of ‘high-level’ nuclear waste has run into similar trouble, and few have found a way forward. Bitter fights with concerned citizens have derailed plans in Germany, Canada and the United Kingdom. “All of the world’s programmes now recognize that the non-technical problems are bigger than the technical ones,” says Charles McCombie, a nuclear-waste consultant who has worked extensively on the Swiss, Canadian and Japanese disposal programmes.

Nations around the world with stocks of waste — including those anticipating rapid growth in such stocks (see ‘Booming nations go nuclear’, overleaf) — are realizing that to dispose of their nuclear waste, they need to build trust with local communities, and that it is on these efforts that much of the future of nuclear power depends. “If we don’t solve the waste

problem, we’re going to have major trouble continuing the revitalization of nuclear power,” says Tom Isaacs, director for policy, planning and special studies at Lawrence Livermore National Laboratory in California.

A lasting legacy

The decay of the isotope uranium-235 in nuclear fuel rods creates atomic nuclei with a wide range of radioactivities and chemical properties. Some, such as neptunium-237, are present only in small quantities but can remain radioactive for millions of years. Others, including caesium-137 and strontium-90, make the waste a health hazard for decades or centuries, and raise its temperature to above 500 °C. The most problematic is plutonium-239, generated from uranium-238, which when purified can be fashioned into a nuclear weapon.

‘Reprocessing’, in which plutonium-239 and unused uranium-235 are reused as fuels or for weapons, reduces the amount of high-level waste produced. But as yet, it has never proved an economic way of making nuclear fuel. And even if reprocessing were to improve dramatically, or if the use of reactors or beams of particles to transmute some of the isotopes into less noxious ones were to prove practical, generating energy from nuclear fission would still produce

“All the world recognizes that the non-technical problems are bigger than the technical ones.”

— Charles McCombie

Booming nations go nuclear

India and China have very little high-level nuclear waste today, but both countries are anticipating substantial nuclear growth spurts to feed their voracious energy needs.

In China, where the number of reactors is expected to go from 9 to 40 within 15 years, geologists are characterizing eight repository sites in remote Gansu province near the Mongolian border. Early work is now well under way, but it will be a while before a site is

constructed, according to Ju Wang of the Beijing Research Institute of Uranium Geology. Initial surveys will be completed in 2015, a site will be selected by 2030, and construction on a final repository could begin by 2050.

Meanwhile, Indian nuclear engineers are planning a massive nuclear complex that will include spent-fuel reprocessing and so-called breeder reactors. These use natural thorium deposits and

reprocessed nuclear fuel to create more uranium fuel. Such breeder reactor systems are technically complex but could dramatically cut down the waste stream coming from India's nuclear complex.

Nevertheless, the nation is still planning a repository, according to Srikumar Banerjee, director of the Bhabha Atomic Research Centre in Mumbai. Experiments have been carried out in a deep mine in the southern state of Karnataka, and further studies are

being carried out in an operating uranium-ore mine, Banerjee says.

Although the technical programmes are well in hand, both India and China have some work to do to resolve the societal issues surrounding a waste dump, notes Malcolm Gray, a nuclear engineer at the International Atomic Energy Agency in Vienna. "The Indians are looking carefully into the social problems," he says. "But, for now, the Chinese think they'll still be able to sidestep it." **G.B.**

some waste posing a hazard for millennia. And that waste will need to be put somewhere.

Since the late 1950s, the scientific community has more-or-less agreed that the best way to deal with this waste is to put it deep underground, says Malcolm Gray, a nuclear engineer with the waste-technology section of the International Atomic Energy Agency in Vienna, Austria. Providing safety and security for surface facilities, such as the spent-fuel ponds used at nuclear power plants, is costly, especially in the long term. Properly chosen and engineered sites can remain stable for millennia and would not require the modification of international treaties on dumping. "Geological disposal is the only practical solution," Gray says. Sites for such disposal are under

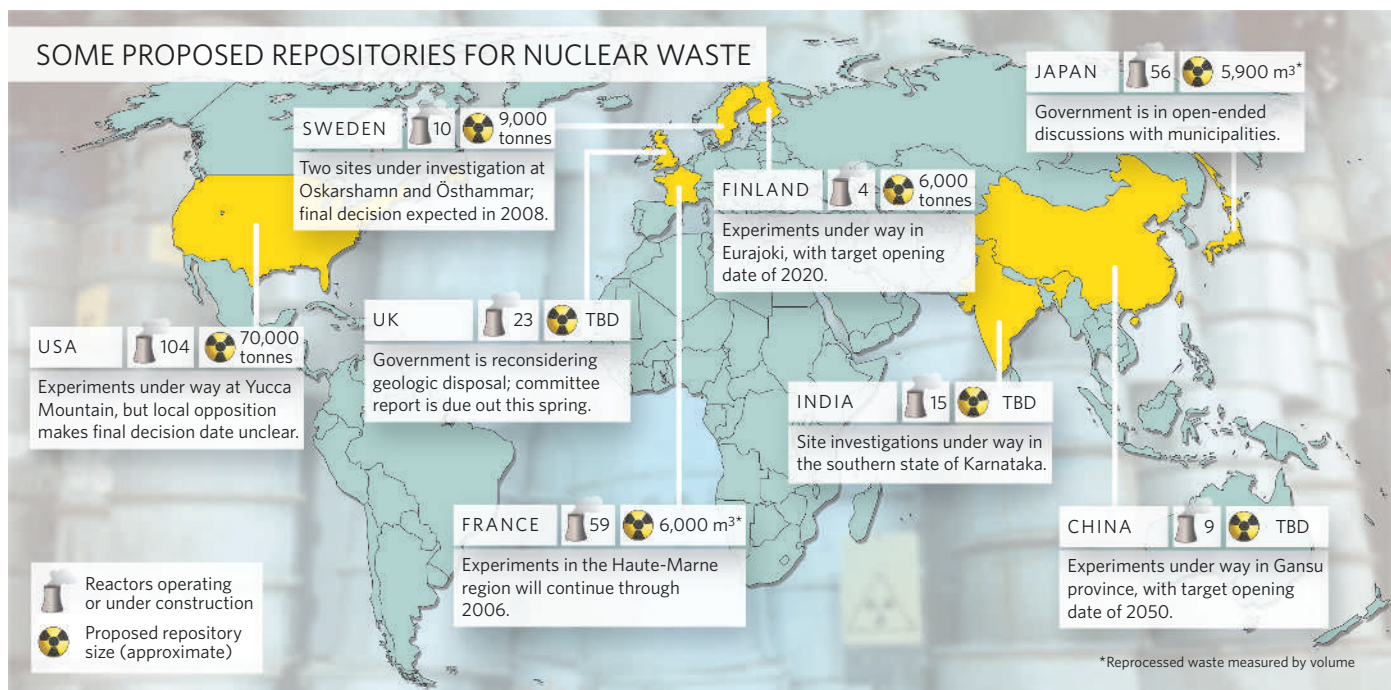
consideration around the world (see graphic).

Yucca Mountain should serve as a cautionary tale, for such efforts, on two grounds. One is that despite years of research it is still not clear how well this particular site would meet the stringent regulations called for by the National Academies of Science (NAS). The NAS recommends a regulatory period of several hundred thousand years, which could mean keeping the waste safe all the way through the next two ice ages. The other is that the local people have never been more united in their rejection of the plans.

Yucca Mountain has been studied as a possible site for nuclear-waste disposal since 1978. A bevy of experiments in an eight-kilometre tunnel built in the side of the mountain have

looked at how water, heat and stray radioactive material might move through the rock. The data are used to estimate exposure rates for humans who might stumble across the site centuries or millennia into the future.

In recent years, some of those experiments have shown that waste could migrate to the water table more quickly than expected. And studies of young volcanoes around the mountain have raised concerns about a breach occurring at the site (see *Nature* 412, 850–852; 2001). A big problem is that no models can be expected to precisely predict exposure over such vast time periods. As Michael Voegelé, a geological engineer with more than 25 years of experience at the site, puts it: "It's impossible to know exactly how water is going to move through the moun-





SKB Keep out: storing radioactive waste in surface facilities such as spent-fuel ponds is expensive and risky.

tain when Chicago's under 5,000 feet of ice."

These uncertainties fuel the doubts of Las Vegas, says Peggy Maze Johnson, executive director of Citizen Alert, an environmental group that is trying to block the project. Local people think Yucca Mountain was picked because Nevada is thinly populated and has little political representation in Washington. "I am not an expert," Johnson says. "All I know is that when something is based on politics, it doesn't make it sound science."

Looking to the leaders

Repositories cannot be built without addressing this blend of political and scientific concerns, says Tero Varjoranta, director of nuclear waste and materials regulation at Säteilyturvakeskus (STUK), a nuclear research centre and regulatory authority in Helsinki, Finland. Varjoranta says it is critical that government agencies be, and be seen to be, impartial and authoritative. "We're not selling the repository," Varjoranta says of his own office. "We try to convince people that if it is built, it will be safe."

In neighbouring Sweden, such impartiality has been vital because of the strength of local municipalities, says Saida Engström, a member of the board of directors at Svensk Kärnbränslehantering AB (SKB). Based in Stockholm, SKB is a company formed by the government that is charged with developing Sweden's waste repository. When Engström was in charge of assessing the environmental impact of a repository in the village of Tierp, 67% of the population supported her work, she says. But the city

council vetoed the project by one vote.

"I'd be lying if I told you I wasn't bothered by it — I was," she says. "But for the credibility of the project it is vital that you involve local people in the process." Ultimately, the veto at Tierp showed that SKB was serious about respecting local wishes and strengthened the programme's reputation in other parts of the country, she argues. The company is continuing investigations at Oskarshamn and Östhammar, where public support for the project continues to be strong.

Attempts to gain public trust have put the two Scandinavian nations far ahead in their efforts to build a repository, and other countries are now looking to them for inspiration. In the United Kingdom, a government committee was established in 2003, to tackle radioactive-waste management. The committee, which includes

"The State of Nevada is committed to stopping this thing by any legal means possible." — Steve Frishman

environmentalists and social scientists as well as physicists and engineers, will determine whether a deeply buried repository is the best way to dispose of waste. Lessons have been learnt after plans to build a repository at Sellafield were thrown out by a public inquiry in 1997. "That was the end of an old tradition where scientists, industry and government got together behind closed doors, thought up the right option, thought up the right site and then

announced it," says Gordon MacKerron, an economist from the University of Sussex who chairs the committee.

Yet others worry that the pendulum has swung too far in the direction of public opinion. "There's a view out there, which basically says that science is not reliable anymore," says David Ball, a professor of risk management at Middlesex University in London. Ball resigned last spring from Britain's committee on radioactive-waste management, in protest of what he saw as an over-reliance on public opinion. "They didn't see science as holding any objective truth and replaced it with the good old common sense of the man on the Clapham omnibus," he says.

Winning trust

Keith Baverstock, a radiation health expert at the University of Kuopio in Finland (see page 993), adds that the Finnish and Swedish success is not due to national public acceptance. "The Finns have never had more than about 40% of the general population agree that this form of disposal was the correct way to do it," he says. The key, he says, is to find a local community willing to accept the project.

The views of local communities will not be determined only by local issues, such as jobs or worries about contamination. "Many groups began opposing Yucca Mountain because they didn't want new nuclear power and they didn't want it seen as the solution that allows for the construction of new nuclear plants," says Judy Treichel, executive director of the Nevada Nuclear Waste Task Force, which opposes the Yucca Mountain project. And in Sweden, there might be more local resistance to the country's waste-repository project if the government wasn't already committed to phasing out nuclear power across the country. At least, that's the view of Johan Swahn, director of the Swedish Office for Nuclear Waste Review — an environmental group that uses government funding to monitor the repository project. "People are more willing to participate in the process with the understanding that nuclear waste is a finite problem," he says.

Regardless of what scientists and engineers believe is technically possible, they must be prepared to address these sorts of cultural concerns, says nuclear-waste consultant McCombie. More broadly, governments eager to see a nuclear repository built must work harder to win their citizens' respect. That's no mean feat and can only be achieved through time and transparency, McCombie says: "You can't just build it like you would a filling station; it has to be a long process with buy-in and the potential for reversal at each stage."

Geoff Brumfiel is Nature's Washington physical sciences correspondent.

BUSINESS

Handle with care

China's drug market is booming, and will soon be the fifth largest in the world. But despite this potential, global companies are being advised to approach the country with caution, as Colin Macilwain reports.

On the face of it, China is a mouth-watering proposition for international drug companies. In the past five years, the value of its pharmaceutical market has doubled to US\$13 billion, according to the Boston Consulting Group — and it is set to double again by 2010. That would make it the fifth largest drug market in the world. With millions more hospital patients than any other nation and an exploding number of life-sciences graduates, basic and clinical research in China should be poised for take off.

Yet last month, at a meeting in London to discuss opportunities in China, the mood was decidedly cautious. Drug-industry managers were advised to proceed into China with care.

Although all of the global players have established sales and marketing operations in China, it is only in the past few years that they have begun sizing up the opportunities for research and development (R&D). For foreign businesses, establishing R&D capabilities in China could be the key to success — it would demonstrate commitment and offer 'added value' to the domestic market, elements that should help break down barriers and win product acceptance.

As a result, most of the major firms have set up research labs in China. But in global terms, these commitments remain "a drop in the ocean", says Simon Friend, head of the global pharmaceutical group at PricewaterhouseCoopers (PWC), the management consultants that co-organized the London meeting. "There's a great deal of nervousness among chief executives" about investing resources in China, he says.

Drug companies still find it tough to make any money in a country where, according to PWC, the government has cut drug prices no fewer than 17 times since 1997. Demand for medicines is strong: but the bulk of them are supplied, at moderate cost, by a highly fragmented domestic industry of some 3,700 certified manufacturers. Twenty-five years after the first signs that China was opening up to foreign suppliers, notes Kenneth DeWoskin a PWC consultant based in Shanghai, 97% of drug demand is still met by locally produced generic drugs — or by counterfeits.

"Prospects are good, but it's a question of finding the right recipe."



Get real: the bulk of Chinese drug demand is met by locally produced generics and counterfeits.

And the outlook remains clouded by uncertainty over China's attitude to intellectual property, DeWoskin argues. The big global firms make almost all of their profits from patented medicines but they just can't assume that China will adapt a model of patent enforcement that will protect profits, he says. "We don't believe that China is on a path that will create an intellectual-property environment that is at all familiar," he explains. Instead of seeing intellectual property as a way for companies to add value, the Chinese "think that low-cost intellectual property will drive Chinese economic growth".

For an example of what might be coming, DeWoskin suggests that drug-makers look at what has happened in other industry sectors. In the production of DVD players, he says, Chinese manufacturers have driven down the price that they pay for the patented ideas from \$20 to \$1 per unit.

And in transgenic crops, China has not given regulatory approval for imported crop strains, to allow time for work on the strains on which it owns the patents to catch up (see *Nature* 422, 111–112; 2003). In China,

DeWoskin notes, "research and development is used to drive intellectual property costs down".

This view is disputed by some other analysts. Ian Harvey, chairman of the London-based Intellectual Property Institute, says that patents are being protected with increasing success in Chinese courts (see *Nature* 438, 420–421; 2005). "All the signs are that China wants to have a strong intellectual-property system," he argues.

Land of opportunity

The largest strand of activity for most global firms so far has been the outsourcing of animal testing and clinical trials. Drug industry officials at the London seminar hinted that light regulation can sometimes be an enticement in this regard. "It is a unique position with regard to the regulatory environment," enthused Sundee Lal, director of external scientific affairs at Merck. "There are tremendous opportunities in animal research." He notes that Merck has now been working in China since it began collaborating on AIDS research there in 1989.

Yet drug-industry leaders know that if they are to establish a real foothold in China, they need to do more, in research terms, than subcontract out animal and clinical trials. In a

report produced to accompany the seminar, PWC recommends joint ventures with indigenous drug makers but, in the absence of any real “national champions” to form ventures with, notes that companies “face a formidable learning curve on the way to building a successful business”.

Full drug discovery inside China remains the goal for some. Andreas Tschirky, head of the Roche lab that opened 18 months ago in Shanghai, claims that goal is not far away. “Within the next few years we’ll have the whole drug-discovery set-up there,” he says. But others are unsure how such arrangements can retain staff, or how they will make an original contribution to global drug development.

“Prospects are good, but it’s a question of finding the right recipe,” says John Stageman, vice-president for emerging projects at AstraZeneca, which is currently crafting its Chinese drug-discovery strategy. “All of the major pharmaceutical companies are struggling with that.” AstraZeneca has 2,000 people in China now, and is the most successful company in terms of prescription drug sales. It has 120 people in R&D — not many, as Stageman admits. But he is convinced that the human talent and enthusiasm for science in China must present opportunities. “There’s a very strong medical culture, with established routes to patients,” he points out. “And — in contrast with the West — the brightest young people in China want to do science.”

Technical strengths exist, but are patchy. According to Rachel Lee of Boston Consulting, China has strong capabilities in some areas of early drug research. But, with the exception of new areas such as gene-sequencing and bioinformatics, biology is not well developed, she says. Chemistry is very strong: researchers can find and synthesize drug candidates, and advanced techniques such as computer-aided drug design are gaining ground.

Yet innovation remains rare in the domestic industry. Despite talk of an emerging biotechnology sector, it is barely extant. Shanghai Genomics, which was founded in 2001 by a group of émigré scientists returning from the United States, is typical. It follows a common business model: it does contract research for foreign partners to build revenues to help it develop its own drug candidates. “Our strength is that we know how to work with Western companies,” says co-founder Jun Wu. “We rely on revenue from that for survival and growth.”

With all these factors confronting foreign firms, Friend warns there will be no easy pickings. “None of the issues is going to be solved overnight,” he says. “The risks are high and those entering China need to do so with their eyes wide open.” ■

IN BRIEF

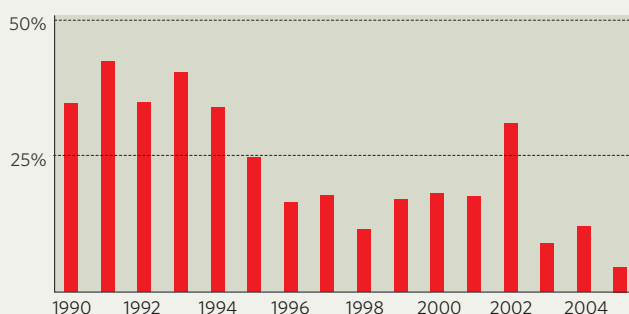
OFF THE BLOCK Serono, Europe’s largest biotechnology company, is no longer up for sale. Instead, the firm says that it is seeking allies and smaller acquisitions to strengthen its drug pipeline. Shares in Serono, whose main product is the multiple-sclerosis drug Rebif, fell back 10% on the news to Sfr830 (US\$640), down sharply from Sfr1,090 in January. The Geneva-based company said last November that the investment bank Goldman Sachs was seeking a buyer or partner for it (see *Nature* 438, 557; 2005) — but none came forward with the right offer.

JOINING FORCES DuPont and Syngenta are teaming up to cross-license each other’s transgenic-crop technology. Delaware-based DuPont, through its subsidiary Pioneer Hi-Bred, will join one of Swiss firm Syngenta’s existing ventures to form GreenLeaf Genetics. This will license out both companies’ technologies to other companies that sell seed directly on to farmers. The combined operation will square off against Monsanto, which is the market leader in genetically modified corn and soya bean seeds.

CYBER SPACE Britain has awarded QinetiQ a £1.8-million (US\$3.2-million) contract to build a security network to help the government, police, banks and computer companies keep an eye on cyber crime. The Cyber Security Knowledge Transfer Network will allow the government and the private sector to share up-to-date solutions to computer security problems. QinetiQ, which runs many of Britain’s former military research laboratories, was floated on the stock market in February (see *Nature* 439, 911; 2005).

MARKET WATCH

NET LOSS/REVENUES, US BIOTECHNOLOGY INDUSTRY



After 30 years of red ink, the biotechnology industry is growing acquainted with a new concept: breaking even.

Last year, according to an annual report released by accounting firm Ernst & Young, the net losses of publicly traded US biotechnology companies fell by more than half, reaching a historic low of 4% of total revenues.

“We’re very optimistic about the global industry being profitable by the end of the decade because of the significant decrease in the net loss this year — and the continued growth of the industry,” says Richard Mejia, a life-sciences partner at Ernst & Young.

A sharp drop in net losses, from US\$4.9 billion in 2004 to \$2.1 billion in 2005, doesn’t achieve a record for the smallest losses in absolute terms: they were lower several times in the

1990s. But relative to revenues, the drop is significant. It’s “a marker of the maturation of the industry”, says Geoffrey Porges, a biotechnology analyst at Sanford C. Bernstein in New York.

The Ernst & Young report notes that it has an in-built bias in its tracking of profitability: it doesn’t include the profits of biotech companies that have been bought out over the years by big drug firms. “If the profits generated by those companies were included, the sector may well have reached ‘profitability’ years ago,” says Michael Hildreth, leader of Ernst & Young’s biotechnology group.

On the other hand, the figures don’t include privately held companies, which tend to be earlier-stage and deeper in debt, and so might swing the equation the other way.

Meredith Wadman ■

Ethics of using employees' eggs in cloning research

SIR — The Hwang case highlights issues in human egg donation that were not addressed in your Editorial “Standards for papers on cloning” (*Nature* **439**, 243; 2006). Developing clones with eggs obtained from one's employees raises serious ethical concerns (see D. Magnus and M. K. Cho *Science* **308**, 1747–1748; 2005).

First, there is considerable risk that the decision to donate is made under pressure and is not entirely voluntary. Second, donors may not be adequately informed. For example, if donated eggs are sought purely for research purposes, the donor must know that they will not be used to develop therapies. Scientists, like all professionals, have an ethical imperative to serve certain socially valued goals, but they must not violate others' autonomy in the pursuit of those goals.

Human eggs are not easily obtained: the process involves trips to a clinic, ultrasound scans, injections to stimulate egg production and, when appropriate, having a probe with an attached needle guided by ultrasound inserted through the vaginal wall into the ovary to remove the eggs. Researchers using human eggs should be independent of any fertility clinics treating the women from whom the eggs came. That way, women are less likely to feel coerced into donating their eggs, and it helps ensure that clinical decisions are not motivated by either scientific or financial gain in the pursuit of these unique stem-cell lines. Cloning publications should include clear information about the steps taken to ensure that egg donors gave their informed and voluntary consent to donation.

Susan Hawes*, **Justin Oakley†**

*Laboratory of Embryonic Stem Cell Biology, Monash Institute of Medical Research, Monash University, Victoria 3800, Australia

†Centre for Human Bioethics, Monash University, Victoria 3800, Australia

Ethics: China already has clear stem-cell guidelines

SIR — As scientists and ethicists who care about stem-cell research in China, we disagree with the statement in your News story “Panel clarifies stem-cell rules” (*Nature* **440**, 9; 2006) that “China lacks clear national policies, with different institutes following different rules”.

In fact, China's government has issued several guidelines to regulate human stem-cell research. These include guidelines on human assisted-reproductive technologies, issued by the Ministry of Health in July 2003, and ethical guidelines for research on human embryonic stem cells, jointly issued by the Ministry of Science and Technology and the

Ministry of Health in December 2003. Both explicitly prohibit human reproductive cloning, and the latter is similar in principle to the guidelines proposed by the US National Academies (www.nap.edu/books/0309096537/html).

It is true that national policies on human stem-cell research in China are not laws. With some further improvement, however, we think they are adequate, as nearly all scientific research in China relies on government funding. There have been cases in China where a few medical practitioners have used human fetal tissues or cells to treat patients, without required government approvals or appropriate clinical trials. We believe that this practice is against commonly accepted principles of modern scientific research. Infringements are a matter of law enforcement against unapproved medical practices, as in any lawful and civilized country, and should not be viewed as unethical examples of human stem-cell research in China.

Linzao Cheng

Johns Hopkins University School of Medicine, 733 North Broadway, Baltimore, Maryland 21205, USA
lcheng@welch.jhu.edu

Other signatories of this letter:

Ren-Zong Qiu Center for Bioethics, Peking Union Medical College, China

Hongkui Deng Peking University College of Life Sciences, Beijing, China

Yu Alex Zhang Capital University of Medical Sciences, Beijing, China

Ying Jin Institute of Health Sciences, Shanghai, China

Lingsong Li Peking University Stem Cell Center, Beijing, China

Eastern European science needs sweeping changes

SIR — Your News story “Ukraine scientists grow impatient for change” (*Nature* **440**, 132–133; 2006) touches on the situation and potential growth of scientific research in a single country, but the issues are relevant to all of the former Soviet bloc.

The facts are sobering. Although the average gross national product per capita in these countries is only a few times lower than in the rest of Europe, the average university ranking is an order of magnitude poorer: in the latest Academic Ranking of World Universities, only 4 of the top 123 European universities are from the former Soviet bloc. Pumping extra money into the system would make little difference. As a member of the Independent Academic Forum (www.nauka-educacja.tubaza.pl) — a group of Polish scientists aiming to promote changes in higher education, leading to the US model — I believe the only real hope lies in creating a new generation of dynamic scientists to set the pace for academic life, which means supporting the best of the best. But much of the old guard, who attained their positions and influence under the old regime, are not up to the scientific challenges of today and resist any real change.

I believe that what we need is transparency and competitiveness: transparent records of achievements, including full publication lists, and fair and open competition for academic positions. An academic ombudsman would facilitate open discussion, and special grants for young scientists could also help. Major political and legislative decisions are called for. The Independent Academic Forum is doing its best to press for such changes.

Witold F. Palosz

BAE Systems Analytical Solutions, NASA, Marshall Space Flight Center, VP01, Huntsville, Alabama 35812, USA

Reviewers peering from under a pile of 'omics' data

SIR — An increasing problem for reviewers, in providing adequate reviews for science journals, is not simply fraudulent data submission or manipulation (see Correspondence *Nature* **439**, 782–784; 2006), but the information density and sheer bulk of data that now have to be supplied as part of publishing modern biological science. This is particularly true with ‘omics’-type data sets (transcriptomics, proteomics, metabonomics and so on), which are now collected in parallel in systems-biology studies.

Many referees are experienced and learned scientists, but they are also very busy people who may well get several papers a week to referee. Do we really have time to read the 60-plus pages of supplementary data that often accompany a major paper? Do we even have the tools and expertise needed to analyse and check the veracity of raw ‘omics’ data sets? A typical data set formatted to meet MIAME (minimum information about a microarray experiment) requirements may contain millions of discrete data.

To check whether these data have been scaled, normalized and processed correctly — within a data set that might have taken a couple of postdocs two years to process — is a difficult task, even if the referee has the time, the knowledge and the right software.

In the data-rich ‘omics world’ of today, the referee's task has become more complex and challenging than could have been envisaged only a few years ago.

Furthermore, there is increasing demand for integrative papers that cover many types of bioanalytical measurement and multivariate statistics at different levels of biomolecular organization. The scientific community needs to reassess the way it addresses the peer-review problem, taking into account that referees are only human and are now being asked to do a superhuman task on a near-daily basis.

Jeremy K. Nicholson

Biological Chemistry, Biomedical Sciences Division, Imperial College London, London SW7 2AZ, UK

COMMENTARY

Too soon for a final diagnosis

Twenty years ago, the nuclear accident at Chernobyl exposed hundreds of thousands of people to radioactive fallout. We still have much to learn about its consequences, argue **Dillwyn Williams** and **Keith Baverstock**.

Why should we still be concerned about the Chernobyl accident after 20 years, some 3,000 papers and many conferences? One reason is that the consequences of the world's worst peacetime nuclear accident are relevant to national debates about building new nuclear power stations. Another is that at this point we cannot predict the future health consequences of Chernobyl with any certainty.

The radiation exposure from the Chernobyl accident differed greatly from that created by the atomic bombs in Japan, yet only fragmented studies have tracked the human consequences of Chernobyl, in contrast to the coordinated approach of Japan's Atomic Bomb Casualty Commission (now the Radiation Effects Research Foundation). Without a similar approach, speculation about Chernobyl's human cost will be unconstrained by hard evidence, and interested parties will be able to exaggerate or underplay the consequences.

This month, UN agencies will mark 20 years since Chernobyl by publishing an international

assessment of the accident's health, environmental and economic effects. A draft¹ issued last year detailed the health consequences and predictions for the future, but the accompanying press release downplayed the predicted number of cancers, emphasizing reassurance rather than the uncertainties.

Misplaced confidence

For example, simply by citing a specific number — up to 4,000 predicted deaths from radiation exposure — the press release suggested a certainty unwarranted by the underlying studies (see Special Report, page 982). The figures quoted in the body of the report suggested that there may be an additional 5,000 radiation-related deaths in heavily contaminated regions. Yet these predictions ignore the large number of Europeans who received very low radiation doses. The dose-response relationship at low doses remains uncertain; it could be linear, but also higher or lower^{2,3}. If it is linear, there may be tens of thousands more attributable deaths.

In 1965, 20 years after the atomic bombings in Japan, the Atomic Bomb Casualty Commission reported significant increases in the incidence of just two cancers — thyroid cancer and leukaemia. It was another decade before a significant increase in other cancers was reported. Almost 45 years after the bombs, unexpected and significant increases in a range of non-cancer diseases, including heart disease, were found⁴. And nearly 50 years after exposure, significant increases were reported for ten different cancers, with risks approximately doubled for colon, lung, breast, ovary and bladder tumours⁵. Given that about one in four people anywhere develops cancer, detecting these and more modest increases in other cancers would have been impossible without the long-term, large-scale studies conducted on some 80,000 Japanese bomb survivors.

In Chernobyl, the type of radiation exposure was different from that in Nagasaki and Hiroshima, and the aftermath will take a different



Check up: a Russian woman living near Chernobyl is examined for signs of thyroid cancer.

L. TCHALENKO/CAMERA PRESS

form. The atomic bombs resulted largely in whole-body radiation from γ -rays and neutrons, exposing all tissues uniformly. Exposure from Chernobyl was, apart from in those working near the reactor, largely internal, from radioactive isotopes in fallout — so different tissues initially received different doses.

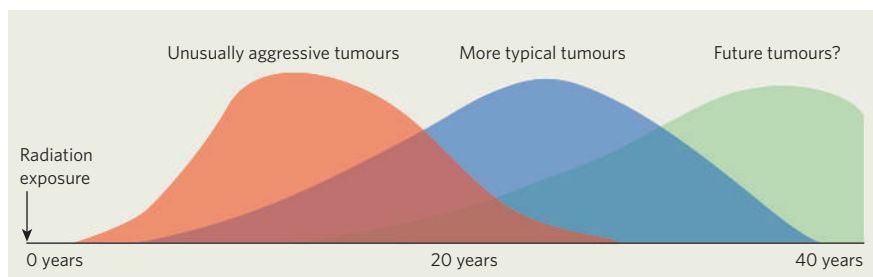
The large increase in childhood thyroid cancer so soon after the Chernobyl accident^{6,7} — the first cases appeared after just four years — took experts by surprise. Experts' expectations were based on known adult exposure risks (minimal) and assumptions about latency times (ten years or more). They were wrong, and there is every reason to expect more surprises over the next 20 years. The high incidence of thyroid cancers has been linked to the huge amount of ¹³¹-iodine in fallout, and to consuming local milk in particular. But the thyroid is not the only tissue to concentrate radioiodine: the salivary glands, breast and stomach also take it up to a lesser extent, potentially increasing their risk of malignancy. Any female who was lactating or pubescent at the time of the accident could be particularly at risk of breast cancer. The future is uncertain, although there is recent evidence for an approximate doubling in breast-cancer incidence in Gomel, Belarus, and other heavily contaminated areas of Belarus and Ukraine⁸. Studies of the atomic bomb show that 20 years is far too early to assume, as some have, that radioactive fallout causes only thyroid cancer.

The most detailed studies since Chernobyl, of the thyroid-cancer epidemic, yielded much valuable information⁹. We have learnt that the risk of developing thyroid cancers depends heavily on the age of exposure to fallout, with children under one being the most susceptible. Several thousand excess cases of thyroid cancer have occurred so far in the three most exposed countries, allowing researchers to link the speed of tumour development with molecular changes and with different tumour types (see figure). The early tumours were clinically aggressive and pathologically unusual. Later ones were more typical and less aggressive, with a changing pattern of oncogene mutation. But the future epidemiology even for thyroid tumours is unpredictable. Other types of thyroid tumour may emerge, and an increase in the lethality of these usually curable tumours cannot be ruled out. Continuous medical surveillance is necessary.

As well as radioactive iodine, Chernobyl exposed millions of people to much lower doses to the whole body from ¹³⁴- and

¹³⁷-caesium. The dose range overlaps with that received by survivors 1 to 2 kilometres from the hypocentre of the atomic bombs. In addition, the ⁹⁰-strontium released is a potential source of bone cancers.

The UN report acknowledges the need for further epidemiological studies to assess the effects of low-dose exposure after Chernobyl; such studies must include verification of the primary data. But existing patchy and uncoordinated studies will not answer all the questions. We believe that, even now, a coordinated



Early cases of thyroid tumours after Chernobyl differed from later ones in unexpected ways.

approach to monitoring exposed populations would at least provide an upper limit to the long-term health risks from fallout.

Fortunately, very few of those who contracted thyroid cancer have died from the disease (15 children so far), but it is too early to conclude that present or future cases will show a similarly low death rate. Other cancers are not so easily treatable. The need for an international effort to monitor all possible health consequences, using the studies of Japanese bomb survivors as a model, cannot be overstressed. At the very least, comprehensive studies should be conducted on the hundreds of thousands in the most affected areas of Belarus, Ukraine and Russia.

A loss of trust

Who should fund such studies? The 2005 budget for the Radiation Effects Research Foundation was US\$40 million. Currently, the United States and the European Union are committed to providing more than a billion dollars to

"Studying the effects of Chernobyl would cost a tiny fraction of the amount spent by the nuclear industry."

make the Chernobyl sarcophagus safe, in addition to the many millions they have already contributed to humanitarian assistance and other studies. We do not suggest diverting funds from these worthwhile efforts, but comprehensive studies of the health consequences of Chernobyl would cost only a tiny fraction of the amount spent annually by the nuclear industry on energy production.

In terms of public health, the psychological consequences of exposure to the accident are probably more important than the physical consequences. Millions were exposed to fallout; all must have some concern for themselves and their children. For hundreds of thousands, fear of the unknown was

compounded by forced evacuation and loss of trust in government, caused by poor risk communication. Those living near the 30-kilometre exclusion zone were troubled by rulings that it was dangerous to live just inside the boundary, but perfectly safe to live just outside.

Public mistrust of the authorities was heightened by mismanaged responses to the accident. At the time, the Soviet nuclear industry and government failed to alert the public to take safety precautions. Later, various national and international organizations downplayed the

effects, while others exaggerated them to gain financial support, often contrary to the best evidence. Environmental organizations and the media have been important whistle-blowers, but have also used unsubstantiated figures, selective quotations and horrific images to paint the

worst possible picture of the consequences. The nuclear industry has been equally selective in its use of figures and quotations, often equating a lack of evidence for an effect with evidence of its absence. As a result, rational public debate is very difficult.

The importance of Chernobyl lies therefore not only in numbers of cancers or deaths, or the economic costs, but also in its effect on people's attitudes. Villagers meeting a team from the International Atomic Energy Agency in 1990 voiced their fear of radiation, fear for their children, and mistrust of officialdom¹⁰: "What chance do the inhabitants of strict control regions have to raise normal children?" "People have been deceiving us for five years — will you tell us the truth?"

If a full, independent study of the consequences of the world's worst nuclear accident is not established, and its results published for all to assess, wildly differing claims will continue, and public mistrust of the nuclear industry will grow further.

Dillwyn Williams is in the Thyroid Carcinogenesis Research Group, Strangeways Research Laboratories, Worts Causeway, Cambridge CB1 8RN, UK.

Keith Baverstock is in the Department of Environmental Science, University of Kuopio, 70211 Kuopio, Finland.

1. International Atomic Energy Authority, Vienna www.iaea.org/NewsCenter/Focus/Chernobyl/index.shtml
2. Brenner, D. J. *et al. Proc. Natl Acad. Sci. USA* **100**, 13761–13766 (2003).
3. Brenner, D. J. & Sachs, R. K. *Radiat. Environ. Biophys.* **44**, 253–256 (2006).
4. Shimizu, Y. *et al. Radiat. Res.* **130**, 249–266 (1992).
5. Thompson, D. E. *et al. Radiat. Res.* **102**, 517–567 (1994).
6. Kazakov, V. S. *et al. Nature* **359**, 21 (1992).
7. Baverstock, K. *et al. Nature* **359**, 21–22 (1992).
8. Pukkala, E. *et al. Int. J. Cancer* doi:10.1002/ijc.21885 (2006).
9. Williams, D. *Nature Rev. Cancer* **2**, 543–549 (2002).
10. The International Chernobyl Project *Tech. Rep.* (International Atomic Energy Agency, Vienna, 1991).

BOOKS & ARTS

Hollow centre

Nanotechnology is a discipline in the throes of an existential crisis.

The Dance of Molecules: How Nanotechnology is Changing our Lives

by Ted Sargent

Thunder's Mouth Press: 2006. 304 pp. \$25

Richard A. L. Jones

Every field needs its founding genius. For many in nanotechnology, this figure is Richard Feynman, on the strength of his 1959 lecture "There's plenty of room at the bottom", in which he discussed the problem of manipulating and controlling things on a small scale. Yet this canonization is entirely retrospective: there is little evidence that Feynman's lecture made much impact at the time, and he rarely returned to the topic to develop his thoughts.

Perhaps a better candidate for nanotechnology's father figure is former US president Bill Clinton, whose support for the National Nanotechnology Initiative converted overnight many industrious physicists, chemists and materials scientists into nanotechnologists. In this cynical — although popular — view, the idea of nanotechnology did not emerge naturally from its parent disciplines, but was imposed on the scientific community from outside. As a consequence, nanotechnology is a subject with an existential crisis — is there actually any firm core to this subject, any consensus as to what defines nanotechnology?

This is the problematic territory that Ted Sargent tries to map out for general readers in *The Dance of Molecules*. For example, he defines the goal of nanotechnology as "to design and build matter to order, specified by a functional requirement". Fine, but it may leave followers of an earlier discipline, materials science, puzzled, as this is their slogan too. He begins by maintaining the centrality of quantum mechanics, but really this is just an assertion of the centrality of chemistry. The book's title might point to brownian motion, but this idea isn't pursued. Sargent is forced to conclude that nanotechnology's central theme isn't scientific, but sociological — a culture of interdisciplinarity that searches for convergence between increasingly atomized scientific fields.

There is one version of nanotechnology that does have clarity: Eric Drexler's vision of scaled-down mechanical engineering. It is this revolutionary vision that underlies much of the popular image of nanotechnology as seen in science fiction, films and computer games. Yet very few scientists take this view seriously. This



J. BOURG/REUTERS

Clear target: the Institute for Soldier Nanotechnologies applies science to meet military needs.

leaves a problem for popularizers who wish to reflect the scientific consensus. They can rebut the Drexler vision in detail, or simply dismiss it with appeals to the authority of scientists such as the late Richard Smalley. Sargent takes a third course: he simply does not mention it. This seems to me to be the most unsatisfactory approach of all. If one thinks the Drexler vision is wrong, one should say so, otherwise the reading public will be confused.

Lacking a strong core of science, the book is written thematically, as a tour of applications to health, environment and information. Quantum dots appear frequently, and the description of molecular electronics is duly circumspect about the balance of potential and practical difficulty. The descriptions of bio-nanotechnology carry less conviction — and that of molecular motors is misleading. Many will find Sargent's overwrought style irritating; perhaps the oddest of the many strange and strained metaphors and similes is his description of photolithography as being "like crop circles formed when the sun blazes through round partings in the English permacloud".

Nanotechnology, above all an applied science, has experienced a possibly unprecedented push for an early consideration of its social, environmental and ethical impacts. Here, Sargent's rhetoric is overwhelmingly positive, and the need for public engagement is seen solely in terms of defusing possible

opposition. We're promised an end to cancer, sight for the blind, and, via unconventional solar cells and the hydrogen economy, an end to our dependence on fossil fuels. The possible downsides are largely limited to the potential toxicity of some nanoparticles. Even in military applications, the emphasis is on defensive applications and on the possibility that nanotechnology will make it much easier for the West to wage a 'clean war', in which combatants are easily distinguished from everyone else. I don't think you need to be a radical anti-technology activist to greet this claim with some scepticism.

The difficulties for nanotechnology include its incompletely formed disciplinary identity and lack of clear definition, the overselling of its immediate potential economic and societal impacts, and its association with extreme utopian and dystopian visions. A good popular book could help to overcome these difficulties by setting out the science that underpins nanotechnology, making realistic claims for what applications and impacts will be possible and when, and presenting a more sophisticated understanding of the relationships between science, the economy and society. This book does not fulfill this need. ■

Richard A. L. Jones is in the Department of Physics and Astronomy, Hicks Building, University of Sheffield, Hounsfield Road, Sheffield S3 7RH, UK.

Looking for clues

Murderous Methods: Using Forensic Science to Solve Lethal Crimes

by Mark Benecke

Columbia University Press: 2005. 304 pp.
\$24.95

Anthony Busuttill

Judging by what we see in the media, the public seems to have an insatiable appetite for reading about murder. This is especially true for serial murders or ones that are particularly heinous and perverse, or where attempts were made to conceal the crime. People like to exercise their 'little grey cells' and try to solve the crimes themselves.

In *Murderous Methods*, Mark Benecke refers to a multitude of homicide cases, some more recent than others. He explains how the evidence was collected and critically discusses how the case came to be solved.

Benecke's pedigree is impeccable: he has a vast knowledge of forensic science, involving such activities as facial reconstruction, forensic anthropology, thanatology (the description and study of death and dying) and fingerprint comparison. He has been involved personally in, or has carefully studied, some *causes célèbres*, both in his native Germany and worldwide, including the Lindbergh kidnapping and the O. J. Simpson case.

This strange mélange of cases is an unorthodox contribution to the literature on forensic science. I suspect that the book is really aimed towards the interested and experienced lay reader, rather than a scientific audience, given the author's close association with the media and his failure to delve to any significant depth into the scientific background of the investigation.

The book is anecdotal rather than didactic. It does not address any particular theme, and the five chapters do not appear to be obviously interlinked. The cases it discusses are all genuine, with no romanticized, intricate or intriguing plots of the sort you would find in a best-selling thriller, but they serve to demonstrate the deviousness and perverse cruelty of human nature. The author also regularly diverts the reader's attention to para-forensic subjects, such as the embalming of Lenin's body, the possibility of life after decapitation, the rate and manner of decomposition of a human cadaver, and so on.

Benecke dispels the myth that all you need to solve crimes is sophisticated DNA profiling, some expensive imaging and magnification equipment, and a team of scientists slaving away in a laboratory. In truth, it often requires meticulous and painstaking police sleuthing, the laborious collection of evidence, careful interviews of witnesses, the application of common sense and principles of logic, and close adherence to the tenet that nothing is quite



Stopping the rot: a combination of chemicals and frost helped to preserve Lenin's corpse.

what it seems. Only when the evidence has been carefully collected, sifted and collated can the case can be presented for judicial scrutiny.

The book explores the judicial systems of mainland Europe, with special emphasis on Germany, which is perhaps the cradle of science-based investigation of crime. Germany has an inquisitorial legal system, in which the judges seek to determine the facts of the case. This is different to the adversarial system used in Britain, and other countries linked to Britain, in which the judge acts as an impartial referee between two opposing parties, with one side attempting to demolish the case presented to the court by the other side. This aspect is

not discussed as openly in other similar books.

Murderous Methods was originally published in German in 2004, and in this English version the narrative is wooden and in places slightly incomprehensible. The few illustrations, although authentic, are poorly reproduced and add little value. Most of the references at the end are not to scientific journals or books, but to general literature, although they are reasonably extensive. Overall, however, the book is a good read, being interesting, varied and eloquent — but it is not for the faint-hearted. ■

Anthony Busuttill is regius professor of forensic medicine at the Medical School, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, UK.

Turning to fraud

Intuition

by Allegra Goodman

Dial Press: 2006. 352 pp. \$25

Jennifer Rohn

The writer Allegra Goodman is not a scientist, but she certainly could have fooled me. Her latest novel, *Intuition*, brings back the sights, sounds and smells of a dozen years at the lab bench, stimulating emotions I had forgotten I'd experienced. In lieu of direct research experience, Goodman consulted several scientists in her family and also shadowed some researchers at the Whitehead Institute in Cambridge to produce an exacting portrait of the rise and fall of a cancer-biology lab.

When flailing postdoc Cliff stumbles on a potentially big result, joint lab head Sandy Glass appropriates it as fodder for a sorely needed new grant. This goes against the better wishes of his scientific partner Marion Mendelssohn,

whose more scrupulous approach is respected but has done the lab few favours. The premature result is coaxed into a *Nature* paper, and fame and funding soon follow. But when fellow postdoc Robin, who also happens to be Cliff's lover, suspects that the finding has been faked and accuses him, the close-knit lab begins to unravel.

The story is set in the 1980s, the same decade as the 'Baltimore affair', in which a researcher, Margot O'Toole, accused her boss, Thereza Imanishi-Kari, of fraud and set off a national furor. Yet the story still resonates today: the recent South Korean stem-cell scandal also involved too-good-to-be-true results, lab whistleblowers and the downfall of a mediagenic scientist, Woo Suk Hwang. But *Intuition* diverts our attention from the bare facts to psychological motives: what might make a successful researcher extend that little bit further into the realm of fudging? You

begin to feel that the unthinkable can become at least explainable.

Cliff's alleged fraud is not of the audacious variety epitomized by physicist Jan Hendrik Schön, whose deception led to the retraction of a string of papers in high-profile journals. It is much more interesting, involving cosmetic acts so unconscious that we never know for certain whether they have even taken place, a circumstance that may prevail in many cases of fraud. Aspects of Cliff's data "were so compelling that in his mind they outweighed everything else. He had sifted out what was significant, and the rest had floated off like chaff." Almost unwittingly, the lab colludes. Infected by Sandy's reckless ebullience, Marion throws caution to the wind: "Doubt had been her scientific ally, the whetstone for her sharpest emotions. Now she struggled against doubt as if it were merely an emotion, and not also a kind of intelligence."

Despite its deft illumination of laboratory

wrongdoings, *Intuition* has a more serious message about modern science as a career: far from fulfilling, it is actually psychologically damaging. Scientific similes subtly reinforce this: a Chinese postdoc wears his nonchalance "like safety glasses"; Sandy shoots a wounding look "like liquid nitrogen"; Robin is hurt to be "treated like hazardous material, to be isolated and manipulated with gloved hands".

And postdoctoral angst? It's all here: the dead-end project bestowed with misguided fanfare by a new mentor; fear of being scooped; fear of having to leave the bench to teach; fear of irreproducible results; fear of running out of funding. As early as the first chapter, Cliff's despair "began to melt and pool inside him" until "he was no longer desperate, but simply demoralized and depressed — emotions entirely accepted, even expected, in the lab". Of another postdoc, we learn that patience, diligence, sarcasm and pessimism "all protected

him from failure and hurt". Science is a "dirty game", a "cosmic joke".

Although the book ends in redemption, the aftertaste is still bitter. The scientists have lists of amusing definitions ("experiment: a series of humiliations"), leave warning notes on refrigerators, and engage in postdoctoral banter; such playful antics are astutely observed and very funny, but in the end cannot mitigate the fundamental darkness. All the negative things described in the book ring true, although their concentration in one lab seems to artificially inflate the difficulties of the profession and blunt its periodic joys. Yet in the end, negative aspects make good reading; any 'lab lit' novel set in a world as technical as molecular biology must dwell on obstacles to keep the pages turning. And turn they most certainly do. ■ Jennifer Rohn resides in London, UK. She is a former molecular virologist and is now the editor of LabLit.com.

The zenith of Islamic science

An exhibition in Britain explores a rich scientific heritage.

Philip Ball

"It is sad to relate that no great invention has come for many hundred years from Muslim countries." Comments such as this, from former Archbishop of Canterbury George Carey in 2004, are part of the subtext for *1001 Inventions*, an exhibition on view at the Manchester Museum of Science and Industry until 4 June, which describes contributions to science and technology that have sprung from the Islamic world.

Mechanical engineer Salim Al-Hassani explains in his accompanying book, also titled *1001 Inventions* (Foundation for Science Technology and Civilisation, 2006), that the exhibition was motivated by the perception of a hiatus in Western science and technology lasting well over a thousand years. While many traditional histories of science jump from Archimedes in the third century BC to Gutenberg in the fifteenth century AD, it was during this intervening period that Islamic culture and science reached its zenith (itself an Arabic astronomical term).

The exhibition is particularly timely given the recent tensions between the Western and Islamic worlds. But there has also been a resurgence of interest among Muslim scholars in the evolution of Islamic belief and its relation to the philosophical and scientific traditions of European culture.

There will be few surprises for scientists



Vaults like this one at Isfahan in Iran may have led to the Gothic rib vault.

among the plethora of inventions and discoveries on display. Common scientific words such as alcohol, algebra and alkali are a constant reminder of the debt that contemporary science owes to the Persian and Arabic scholars of the sixth to the eleventh century AD. But it is a debt that receives little public acknowledgement — Western schools, deplorably, still teach of this being the 'dark ages'. As an educational resource the exhibition is impressive. Indeed, even enthusiasts of science history are likely to find something here they did not know: Ibn al-Jazari's exquisite water clock, perhaps, or Ibn al-Mosuli's cataract operations.

The danger is that by trying to redress the imbalance, the exhibition claims too much and clarifies too little. Bald statements of priority, such as that the Gothic rib vault originated in the mosques of Toledo and

Cordoba, or that the European university stems from the organization of Muslim scholarship, are striking but, without any sense of the historical progression, it is like saying that Charles Babbage invented the Apple Mac. The claim that Jabir ibn Hayyan was the founder of chemistry, and the assumption that this eighth-century Muslim scholar was the author of the entire Jabirian alchemical corpus, ignores a great deal of careful scholarship.

The impression that Islamic science sprang up of its own accord is misleading too — not enough is said about how (or why) it helped to preserve the knowledge of classical and Hellenistic Greece. Nor does the exhibition provide much sense of how this learning found its way to the West, a story of cross-cultural exchange that was all the more remarkable for coinciding with the Crusades.

Perhaps this is too much to expect of an exhibition aimed largely at a younger audience. But it is a shame that it fails to address the crucial question of why Islamic pre-eminence in science ended just before the European Renaissance. That is, after all, the point to which Carey was alluding (if not in the most helpful language), and it is one that Muslim scholars are now keen to debate. It would do the riches of Islamic culture no favours to pretend the question does not exist.

Philip Ball is a consultant editor for *Nature*.

NEUROSCIENCE

Spikes too kinky in the cortex?

Boris Gutkin and G. Bard Ermentrout

The Hodgkin–Huxley theory that explains the mechanism of how neurons fire forms the cornerstone of computational neuroscience. But something it hasn't predicted is happening in the brain cortex.

Neurons encode and transmit information by generating action potentials, or 'spikes' of voltage that sweep along their membranes (Box 1). It has been 50 years since Hodgkin and Huxley¹ proposed a mechanistic model by which such spikes are generated in an electrically excitable membrane. The theoretical framework associated with their model, the so-called HH formalism, is the closest that neurophysiologists have to Newton's laws of motion, and it underpins almost all modern models of how neurons work. In a bold and imaginative study, Naundorf *et al.* (page 1060 of this issue)² challenge the HH theory and specifically its applicability to the neurons of the cerebral cortex.

Hodgkin and Huxley¹ showed that spikes in neuronal membranes are produced through the interaction of fast depolarizing and slower hyperpolarizing currents that are dynamically dependent on the voltage (Box 1). Without knowing that voltage-dependent ion channels existed, Hodgkin and Huxley reasoned that the experimentally observed activation and inactivation curves of these currents could be explained by the collective action of a population of channels acting independently. This theory was later backed up with thermodynamical arguments³.

A practical outcome of this framework is a family of computational models stating, in terms of differential equations, how the dynamics of the various voltage-dependent processes evolve with time. In this family of models, the currents are the product of the conductance per ion channel, the density of channels, the fraction of open channels, the membrane area and the electrical driving force due to the ion concentration gradient. One of the key tenets is that the fraction of the open channels is governed by the membrane voltage, with channels themselves being mutually independent.

The HH formalism predicts that, when viewed at high temporal precision, spikes rise smoothly once the membrane voltage reaches the activation threshold. Naundorf *et al.*² have tested this by recording spikes from a variety of cortical neurons *in vitro* and *in vivo*. They found that the spikes rise much more abruptly

than expected from the models based on the HH formalism. Moreover, the voltage at which the spike takes off (the threshold) varies remarkably from spike to spike in the same neuron *in vivo* (Fig. 1, overleaf). Examining several HH-based models, the authors find that none of them can account for both observations — spike sharpness and threshold variability. Indeed, under the HH formalism the spike speed and onset-voltage variability seem to be antagonistic: if the model parameters are tweaked to fit the observations for spike speed, the variability in onset cannot occur and vice versa.

Naundorf *et al.* conclude that the HH formalism is not valid for the neurons that they monitored, and they propose a complementary model that invokes cooperative voltage-gated sodium channels. They took as their basis a sodium-channel model⁴ that allowed for a slow voltage-independent transition from the open state to the inactivated closed state. This can produce the threshold variability: for slower inputs to the neuron the threshold is higher than for rapid inputs, as, during the slow

depolarization phase, more channels are inactivated, so greater input is needed for activation.

Naundorf *et al.* then make a radical addition to this model: they allow the threshold for the voltage-dependent transition from the closed state to the open state to depend directly on the number of open channels. This channel cooperativity means that the opening of one channel makes the neighbouring channels more likely to open, providing a positive-feedback mechanism over and above the intrinsic voltage dependence of the channels. This allows sharp spike onsets without affecting the inactivation rate, leaving the onset variability intact. If channel cooperativity does occur, it should depend on the channel density, with sparser channels producing less cooperativity. So Naundorf *et al.* tested their theory by reducing the effective channel density using the sodium-channel blocker tetrodotoxin. They found that the spikes did indeed become smoother and closer to the HH model.

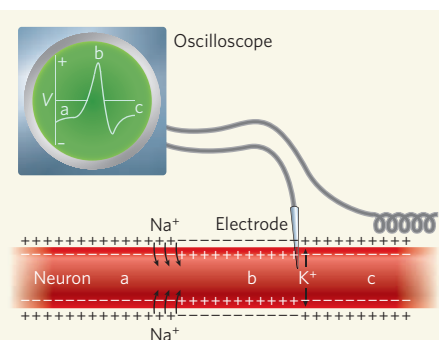
All in all, it's a tall order to contest a theory as well established as the HH model. Even if Naundorf and colleagues' theoretical proposals

Box 1 | Action potentials in neurons

The membranes of all cells have a potential difference across them, as the cell interior is negative with respect to the exterior (a). In neurons, certain stimuli can reduce this potential difference by opening sodium-ion channels in the membrane. For example, neurotransmitters interact specifically with ligand-gated sodium-ion channels. So sodium ions flow into the cell, reducing the voltage across the membrane.

Once the potential difference reaches a threshold voltage, the reduced voltage causes hundreds of voltage-gated sodium channels in that region of the membrane to open briefly. Sodium ions flood into the cell, completely depolarizing the membrane (b). This opens more voltage-gated ion channels in the adjacent membrane, and so a wave of depolarization courses along the cell — the action potential.

As the action potential nears its peak, the sodium channels close, and potassium channels open, allowing ions to flow out of the cell (the hyperpolarizing current) to



restore the normal potential of the membrane (c).

Because action potentials are an all-or-nothing response, occurring only once the threshold voltage is reached, the strength of the stimulating signal will not produce a larger 'spike' in the neuron. Strong stimuli will instead produce a series of action potentials — so it is the frequency, number and timing of the spikes that encode neural information.

B.G. & G.B.E.

are borne out, there are still some difficult questions to be answered. First, what are the implications for the function of cortical neurons — how does this affect the circuits they are involved in? The authors make a start at addressing this question by examining the ability of cortical neurons to respond to various input frequencies; that is, the neurons' filtering properties. They show that sharp-onset, variable-threshold neurons are likely to transmit extremely high-input frequencies while filtering out the very low frequencies — hence such neurons respond much faster than previously expected. This result may help to explain why our neurons have seemed too 'slow' to account for the speed with which we can respond to rapid stimuli (for example, in sensory discrimination experiments). On the other hand, collective behaviours in neural circuits, such as synchronization, may depend on the dynamics of spike generation (for example, see ref. 5); it remains unclear how Naundorf and colleagues' proposed mechanism would affect such large-scale dynamics.

Second, do all cortical neurons display this combination of sharp rise and variable onset? It seems that plots of voltage against the rate of change of voltage are not always as steep as those reported by Naundorf *et al.* (R. Gerkin, T. Bal and Z. Piwkowska, unpublished data). So, if there is a computational reason why some parts of the cortex have this feature, but others do not, then there are lessons to learn about how intrinsic neuronal dynamics affect the circuit properties of cortical networks.

Finally, is channel cooperativity realistic? In cell biology, cooperative behaviour and

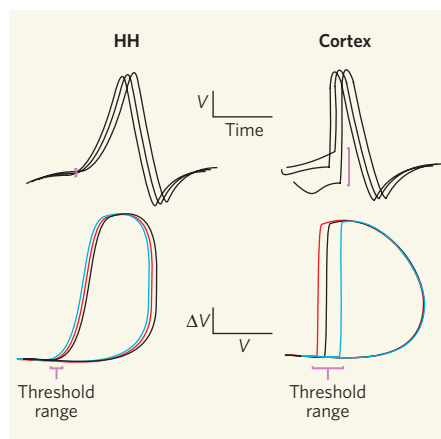


Figure 1 | Deviations from the Hodgkin-Huxley (HH) model in the cortex. The top traces show how voltage (V) changes across a specific point in the membrane, as the action potential sweeps along the neuron. The bottom graphs plot the rate of change of voltage (ΔV) against voltage, so the speed of the spike onset and the threshold voltage can be seen clearly. Under the HH formalism, spikes are smooth with a limited range of thresholds (pink brackets). However, spikes recorded in the cortex by Naundorf *et al.*² are sharp (vertical take-off in the bottom plot) with a wide range of thresholds *in vivo*.

mechanisms of state-dependent self-feedback are quite common. One can even argue that voltage-dependent channels are cooperative at the population level because they depend on the membrane voltage. However, Naundorf *et al.*² propose a rather different level of cooperativity — direct interactions of single channels. Although such cooperativity has

been proposed before⁶, we know of no conclusive evidence that such interactions exist.

But are they likely? A back-of-an-envelope calculation assuming a very high conductance, say $10,000 \text{ mS cm}^{-2}$ (about 100 times the usual value in cortical models), and a single-channel conductance of 10 pS, suggests that individual channels would be about 10 nm apart (with a uniform distribution). Ion channels are only a few nanometres in diameter, so this may be too far apart for any direct interaction. However, there is evidence that ion channels cluster, so there might be some interactions within the clusters. Testing for the existence of such direct and local cooperative interactions will require some ingenuity from the experimental biophysics community. ■

Boris Gutkin is in the Group for Neural Theory, the Department of Cognitive Studies, ENS, and the Receptors and Cognition Department of Neuroscience, Pasteur Institute, 25 rue du Dr Roux, 75015 Paris, France. e-mail: boris.gutkin@ens.fr

G. Bard Ermentrout is in the Department of Mathematics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA. e-mail: bard@math.pitt.edu

1. Hodgkin, A. L. & Huxley, A. F. *J. Physiol. (Lond.)* **117**, 500–544 (1952).
2. Naundorf, B., Wolf, F. & Volgushev, F. *Nature* **440**, 1060–1063 (2006).
3. Borg-Graham, L. in *Cerebral Cortex* Vol. 13 (eds Jones, E., Ulinski, P. & Peters, A.) 19–138 (Plenum, New York, 1999).
4. Aldrich, R. W., Corey, D. P. & Stevens, C. F. *Nature* **306**, 436–441 (1983).
5. Ermentrout, B., Pascal, M. & Gutkin, B. *Neural Comput.* **13**, 1285–1310 (2001).
6. Changeux, J.-P., Thiery, J., Tung, Y. & Kittel, C. *Proc. Natl Acad. Sci. USA* **57**, 335–341 (1967).

GLACIOLOGY

Ice-sheet plumbing in Antarctica

Garry K. C. Clarke

It's not easy to work out what is going on beneath four kilometres of ice. But remote imaging has enabled the discovery of the long-distance discharge of water from one subglacial lake to another in Antarctica.

It is old news that there are lakes under the Antarctic ice sheet¹ — indeed, Earth's seventh largest lake, Lake Vostok², lies deep beneath the Antarctic ice. But are the lakes that form beneath great ice sheets long-lived and stable, collecting and spilling water at a steady rate? Or are they subject to cycles of filling and rapid flushing, like those that form beneath Iceland's ice caps³?

Spectacular landscape features, apparently sculpted by subglacial flood waters, are found in East and West Antarctica^{4,5}, and a flood from an unknown source was once observed on the East Antarctic ice margin⁶. But otherwise, evidence for rapid discharges has been scant — until now, that is, for Wingham *et al.*⁷

(page 1033 of this issue) present convincing observations of such a flood from an Antarctic subglacial lake. The flood transferred water from a small lake near Dome C in East Antarctica to two subglacial lakes situated downslope. As a consequence, the water level in the source lake dropped by 3 metres and those of the receiver lakes rose 1 metre. This subglacial drama, inferred from changes in ice-sheet elevation monitored from a satellite some 750 kilometres above Earth's surface, occurred over a 16-month period and beneath 4 kilometres of ice.

For complicated reasons, the flow of subglacial meltwater is influenced ten times more strongly by the slope of the ice-sheet surface

than by the slope of the underlying bed surface. It therefore requires a very large back-slope to trap water in subglacial ponds or lakes and, as a result, ice sheets tend to expel water rather than store it⁸. Lakes tend to form near the domes of ice sheets (where the surface slope is gentle) or in deep bedrock basins (where the bed slope is steep). The lake featured in the present study is both near a dome and in a bedrock basin.

For all but the smallest lakes, the ice roof is afloat in the lake. As required by Archimedes' principle, the weight of the floating ice is supported by water pressure from the lake. This certainly holds for Lake Vostok⁹, and should be true for other lakes, because ice is too weak to allow it to act as a bridge over any great distance. During discharges, the balance between the downward force exerted by ice and the upward force exerted by water is disturbed, and the unbalanced force causes the ice roof to lower as the lake level drops (Fig. 1). The descending roof can be viewed as an ice piston moving through a cylinder of ice. The motion of this piston is resisted by viscous deformation at the contact between the piston and the cylinder. If the fall in water level overtakes the

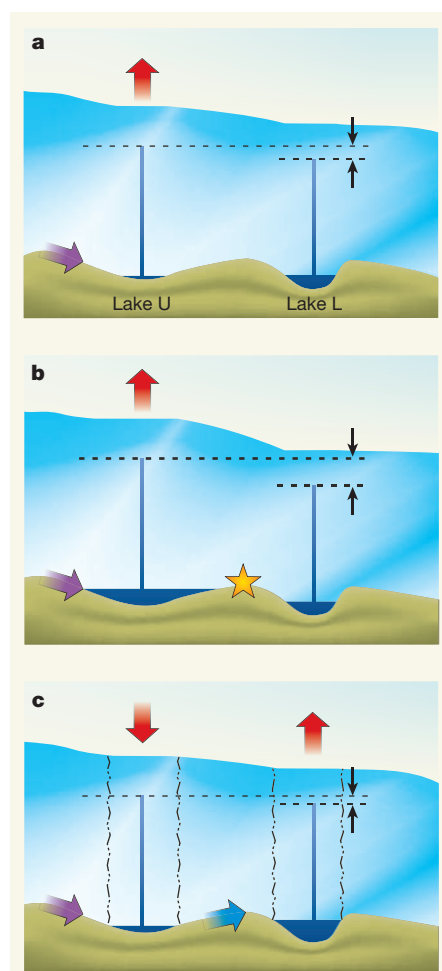


Figure 1 | Cyclic filling and flushing of a subglacial lake system. **a**, Lakes U (upper) and L (lower) are isolated from each other and have differing hydraulic potentials (indicated by the water levels in the hypothetical manometers). Meltwater flows into lake U (purple arrow), raising its level and that of the ice floating on the lake (red arrow). **b**, The level of lake U and the overlying ice column continues to rise, increasing the hydraulic potential difference between lakes U and L (black arrows) until a drainage pathway that connects the lakes develops (star). **c**, Water transfer from lake U to L (blue arrow) causes the level of lake U to fall and that of L to rise. The motion of the ice columns overlying the two lakes can be monitored from space⁷, and can be likened to that of a pair of ice pistons moving through an ice cylinder. Ice deformation at the contact between the piston and cylinder resists this motion, possibly attenuating the flood.

descent of the roof, the ice pressure acting on the lake is reduced, slowing the expulsion of water from the reservoir. This catch-up game could reduce the intensity and increase the duration of some floods.

The characteristics of the observed discharge contrast sharply with those of analogous events beneath Icelandic ice caps. The Antarctic flood path (290 km) and duration are much longer, and the peak flood magnitude is much smaller. Intriguingly, the level in the upper lake dropped only 3 metres before

the flood ended, leaving an elevation offset of 150 metres in the manometer levels (Fig. 1) between the upper and two lower lakes. So why did it stop? It is unlikely that the discharge completely emptied the upper reservoir because, located within a topographic depression in hilly surroundings, it is surely deeper than 3 metres. Nor is the drop in lake level, leading to a slight decrease in water pressure along the drainage channel, sufficient to create a large enough difference between ice pressure and water pressure to pinch off the flow. The best explanation is that details of local topography dictated the outcome, suggesting that floods from other lakes will have individual characteristics.

The discovery by Wingham *et al.*⁷ comes at a pivotal moment in the study of Antarctic subglacial lakes. Plans are afoot to retrieve water samples from Lake Vostok to test the claim that a microbial ecosystem exists in the lake¹⁰. Aspects of this proposal are controversial, but all agree about the need for responsible exploration of such unusual environments. The possibility that subglacial lakes are not isolated entities but are interconnected, or sporadically connected, must now be taken seriously. Accidental contamination of one lake could be spread to others. However, the new information does not rule out the possibility that some subglacial lakes, Lake Vostok for example, are stable and long-lived. For now, the safest bet is that the Antarctic ice sheet hosts lakes of all varieties.

Skirmishes between gradualists and catastrophists have become commonplace in glaciology. Although the new observations⁷

can be expected to embolden the catastrophists, the actual discharge was hardly catastrophic; the estimated maximum rate, around $50 \text{ m}^3 \text{ s}^{-1}$, is small relative to typical floods in Iceland — $5,000 \text{ m}^3 \text{ s}^{-1}$ (ref. 3) — and minuscule compared with the megafloods of the last Ice Age, such as those from glacial Lake Missoula — $10,000,000 \text{ m}^3 \text{ s}^{-1}$ (ref. 11). Other features of the flood are noteworthy, however. Neither the considerable depth of overlying ice, nor the low hydraulic gradient, nor even the great length of the flood path prevented the flood from developing. On these grounds, at least, there is no reason to rule out the occurrence of rapid subglacial discharges of water over much longer pathways both in the past and today.

Garry K. C. Clarke is in the Department of Earth and Ocean Sciences, University of British Columbia, 6339 Stores Road, Vancouver, British Columbia V6T 1Z4, Canada.
e-mail: clarke@eos.ubc.ca

1. Siegert, M. *J. Annu. Rev. Earth Planet. Sci.* **33**, 215–245 (2005).
2. Studinger, M., Bell, R. E. & Tikku, A. A. *Geophys. Res. Lett.* **31**, L12401 (2004).
3. Björnsson, H. *Glob. Planet. Change* **35**, 255–271 (2002).
4. Sawagaki, T. & Hirakawa, K. *Geogr. Ann.* **79A**, 223–238 (1997).
5. Denton, G. E. & Sugden, D. E. *Geogr. Ann.* **87A**, 67–85 (2005).
6. Goodwin, I. D. *J. Glaciol.* **34**, 95–101 (1988).
7. Wingham, D. J., Siegert, M. J., Shepherd, A. & Muir, A. S. *Nature* **440**, 1033–1036 (2006).
8. Clarke, G. K. C., Leverington, D. W., Teller, J. T., Dyke, A. S. & Marshall, S. J. *Quat. Sci. Rev.* **24**, 1533–1541 (2005).
9. Tikku, A. A., Bell, R. E., Studinger, M. & Clarke, G. K. C. *Earth Planet. Sci. Lett.* **227**, 249–261 (2004).
10. Inman, M. *Science* **310**, 611–612 (2005).
11. Benito, G. & O'Connor, J. E. *Geol. Soc. Am. Bull.* **115**, 624–638 (2003).

EVOLUTION

Spot on (and off)

Gregory A. Wray

The repeated appearance and loss of a spot on the wings of fruitflies during their evolution is caused by mutations in one gene. This finding provides an unprecedented window on the genetics of convergent evolution.

Similarities between species can arise in two ways: either each species has retained the trait in question from their common ancestor, or each has acquired it independently¹. Although the first possibility might seem far more likely, convergence is sufficiently common and detailed in its manifestations for more than one unwary biologist to have been duped. Indeed, life abounds with cases of convergence: the wings of birds and bats, and the eyes of octopuses and humans are among the most familiar examples from an immense and varied casebook². But although evolutionary biologists can usually identify independent origins of a trait, the genetic basis of this process is less clear. On page 1050 of this issue, Prud'homme

*et al.*³ describe the identification of mutations responsible for the convergent evolution of a black spot on the wings of various species of fruitfly (Fig. 1, overleaf). The results reveal much about the way in which simple mutations can modify developmental processes to produce convergent traits.

The authors began by reconstructing genealogical relationships among 77 fruitfly species in the genus *Drosophila*. They then used this information to identify gains and losses of a pigment spot on the flies' wings, and found that the spot had evolved independently at least twice, producing convergent similarity in *D. biarmipes* and *D. tristis* (Fig. 1 on page 1050). But the spot was subsequently lost on

several occasions as well. For instance, the unadorned wings of the most famous species in the genus, *D. melanogaster*, derive from a distant spotted ancestor.

Prud'homme *et al.* took advantage of this chequered history to study the genetic basis for convergence. To validate their approach, they examined two spotted species that derive from the same spotted ancestor and whose spots would therefore be expected to share the same genetic basis. Their previous work with one of these species, *D. biarmipes*, identified mutations in a part of the regulatory region of the *yellow* gene (called the spot element) as responsible for one origin of the wing spot⁴. This gene encodes an enzyme involved in pigment synthesis⁵, and mutations affecting its expression have contributed to the evolution of certain pigmentation patterns in other fruitfly species⁶. During wing development in *D. biarmipes* and *D. elegans*, the *yellow* gene is expressed specifically in a region that prefigures the black spot of the adult wing, implying that there is probably a similar underlying genetic basis for the spots in these species.

So much for similarity by common descent; what about convergent similarity? The repeated spot losses and gains turn out to have a surprisingly similar genetic basis. In two independent cases of spot loss, represented by *D. gunungcola* and *D. mimetica*, different sets of mutations in the spot element of the *yellow* regulatory region abolish expression of *yellow* specifically in the spot region of the wing. The spot element comprises just a few hundred of the roughly 180 million base pairs of the fruitfly genome, making this a 'similar' genetic basis by any reasonable criterion.

Even more striking is the case of *D. tristis*, a species that underwent an independent, convergent acquisition of the wing spot. Here, the spot regulatory element does not harbour the mutations responsible for spot production in this species. Instead, Prud'homme *et al.* discovered a different regulatory element that activates *yellow* expression in the spot region of the wing. The convergently similar wing spots of *D. biarmipes* and *D. tristis* are therefore the product of mutations in the same gene, but involve co-option of different regulatory elements.

These results hint at generalities in the genetic basis for anatomical evolution. The spot gains and losses all result from mutations that affect the expression of the gene, reinforcing the notion that such regulatory mutations constitute a major component of the genetic basis for anatomical evolution^{7,8}. The fact that the resulting anatomy is so similar conveys a more subtle and interesting message. The developing wings of fruitflies are patterned by signalling molecules that activate the expression of transcription factors. These proteins in turn activate the spatially restricted expression of genes such as *yellow*, which encode proteins that convert the pattern into anatomical structures. The spatial scaffold for producing a spot

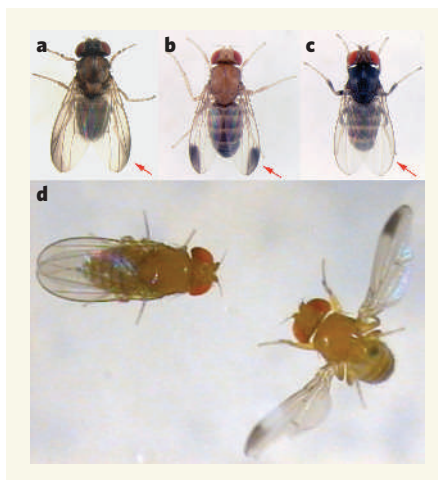


Figure 1 | Spots on flies. **a–c**, The wing spots on male flies of the *Drosophila* genus. *Drosophila tristis* (**a**) and *D. elegans* (**b**) have wing spots that have arisen during convergent evolution. *Drosophila gunungcola* (**c**) instead evolved from a spotted ancestor. **d**, Males wave their wings to display the spots during elaborate courtship dances. (Photographs courtesy of B. Prud'homme and S. Carroll.)

is therefore present even in species lacking spots⁹. If, through random mutation, a gene acquires a new binding site for a transcription factor in its regulatory region¹⁰, its expression is likely to change in ways that reflect the existing spatial scaffold.

This explains similarities in the size, shape and position of independently evolved spots. But why was *yellow* involved each time, given that many fruitfly genes affect pigmentation? The authors argue that a gene already expressed in the wing (such as *yellow*) is much more likely to be recruited to produce a new

wing pattern through a small number of mutations. This is because it already possesses regulatory elements that interact with some of the transcription factors required to produce the new pattern.

Of course, none of this explains why mutations that generate or erase spots become established within fruitfly populations. The reason for that seems to be sex. Male fruitflies woo potential mates using elaborate courtship dances that involve a good deal of wing wagging (Fig. 1d). Accordingly, shifts in mate-choice preferences among females could drive rapid changes in wing patterns^{11,12}. Perhaps because it is already expressed in wings, *yellow* has repeatedly provided genetic variation that produces anatomical differences, making it central to the diversification of fruitfly coloration. ■

Gregory A. Wray is in the Department of Biology and Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina 27708, USA.
e-mail: gwray@duke.edu

- Hall, B. K. (ed.) *Homology: The Hierarchical Basis of Comparative Biology* (Academic, London, 1994).
- Conway Morris, S. *Life's Solution: Inevitable Humans in a Lonely Universe* (Cambridge Univ. Press, 2003).
- Prud'homme, B. *et al.* *Nature* **440**, 1050–1053 (2006).
- Gompel, N. *et al.* *Nature* **433**, 481–487 (2005).
- Wittkopp, P. J., True, J. & Carroll, S. B. *Development* **129**, 1849–1858 (2002).
- Wittkopp, P. J., Vaccaro, K. & Carroll, S. B. *Curr. Biol.* **12**, 1547–1556 (2002).
- King, M.-C. & Wilson, A. *Science* **188**, 107–116 (1975).
- Carroll, S. B. *Endless Forms Most Beautiful* (Norton, New York, 2005).
- Campos-Ortega, J. A. & Hartenstein, V. *The Embryonic Development of Drosophila melanogaster* (Springer, Berlin, 1997).
- Stone, J. & Wray, G. A. *Mol. Biol. Evol.* **18**, 1764–1770 (2001).
- Kopp, A. & True, J. R. *Evol. Dev.* **4**, 278–291 (2002).
- Yeh, S.-D., Lio, S.-R. & True, J. R. *Hereditas* (in the press).

ASTRONOMY

Trouble at first light

Piero Madau

The question of how much light the first stars produced is fundamental to models of the Universe's development. But observations have so far failed to agree: is the answer a lot, or not very much at all?

On page 1018 of this issue, Aharonian *et al.*¹ report the detection of copious high-energy γ -ray emission from two 'blazars' — a class of active galaxy — around 2 billion light years from Earth. This observation indicates that such radiation can travel largely unimpeded through the cosmos, and implies that the infrared glow of the first stars in the Universe and their remnants is fainter than previous measurements had led us to believe. If true, that could influence our ideas of how and when the first structures in the Universe evolved.

The formation of structure in the Universe is believed to proceed hierarchically, with smaller galaxies merging, through the action of gravity, to build more massive ones. But the timing and sequence of the events through which the very first galaxies and stars formed remain largely unknown. According to current theories, the first dwarf galaxies hosted metal-free stars over a hundred times more massive than the Sun. These stars shone intensely for only a few million years and then either blew themselves apart in gigantic supernova explosions, or collapsed

to form the first massive black holes.

Astronomers have long been rummaging through the Universe for tell-tale signs of these dramatic beginnings. When the first stars ignited, they emitted large numbers of photons at ultraviolet wavelengths. These photons 'reionized' the surrounding atomic hydrogen gas that had formed as the Universe cooled. Just last month, astronomers using NASA's Wilkinson Microwave Anisotropy Probe (WMAP) reported the latest detection of photons produced soon after the Big Bang. Their data show that these 'cosmic microwave background' photons became polarized (tending to oscillate in only one direction perpendicular to their line of travel) by scattering on free electrons in the early Universe. The level of polarization allows the era of reionization to be pinpointed to some 400 million years after the Big Bang, when the Universe was just 3% of its present age².

So how much of the background light that we see comes from the first stars? As the Universe aged and expanded, part of the ultraviolet radiation emitted by these stars was absorbed again by re-formed atomic hydrogen. Lower-energy ultraviolet light escaped this fate, but was stretched to longer, redder wavelengths. Therefore, although the early stellar populations were twinkling so long ago that current telescopes cannot detect them, their combined energy output is recorded in diffuse light that reaches Earth in the near-infrared region of the electromagnetic spectrum, at wavelengths of a few micrometres. Resolving this infrared glow is, however, a daunting task, because many other celestial sources — among them older stars in closer galaxies, active galactic nuclei known as quasars, and the bright foreground sources in the Milky Way and the Solar System — emit radiation at similar wavelengths.

Nevertheless, several groups have claimed to have found the footprints of baby galaxies at near-infrared wavelengths, using data from NASA's Cosmic Background Explorer (COBE)³⁻⁵ and Spitzer Space Telescope⁶, and Japan's Infrared Telescope in Space (IRTS)⁷. Their evidence comes in two forms. First, there is an excess signal above the combined emission of normal foreground galaxies that would require energetic events to have occurred in the early Universe. Second, the very uneven distribution of the radiation could arise from the spatial clustering properties of primordial stellar systems.

But rather than helping to decipher the epoch of cosmic first light, such observations have in fact created another puzzle. Simply stated, the dawn of galaxies seems to be too brilliant: the excess signal outshines the cumulative emission from all galaxies between Earth and the extremely distant first stars. If primordial sources are to account for all of this infrared radiation, current models of star formation in the young Universe look distinctly shaky. Too many massive stars ending their



Figure 1 | Eyes to the light. The four HESS telescopes in the veldt of central Namibia.

brief lives in a giant thermonuclear explosion would, for instance, eject large amounts of heavy elements such as carbon and oxygen into space, polluting the cosmos very early on and altering for ever the composition of the raw material available for second-generation stars. But if the first-generation stars were to collapse to massive black holes instead, gas accretion onto such black holes would produce large amounts of X-rays. Both variants seem to be in conflict with current observations⁸⁻¹⁰.

Enter Aharonian and colleagues¹, and their measurements of teraelectronvolt (TeV) γ -ray photons from blazars. These photons, which carry 10^{12} times more energy than visible light, interact with near-infrared photons through the quantum-mechanical process of electron-positron pair creation. Through this process, most of the TeV photons are absorbed long before they reach Earth. The observed level of γ -ray attenuation can, therefore, be used to estimate indirectly the energy density of infrared starlight present in intergalactic space.

The authors' observations were made with the High Energy Stereoscopic System, HESS, inaugurated in Namibia in 2004 and operated by a collaboration of scientists from nine countries. HESS uses four large telescopes (Fig. 1) arranged at the corners of a 120-metre square to detect the faint flashes, lasting only a few billionths of a second, of blue 'Cherenkov' light emitted when a high-energy γ -ray hits the atmosphere. Up to four images are combined to determine the direction of the γ -ray and the energy it deposits in the atmosphere.

Since June 2004, HESS has accumulated more than 80 hours of observations on two blazars. The γ -rays emitted by these most distant known sources have energies between 0.2 and 3 TeV; infrared radiation at wavelengths longer than 1 micrometre absorbs γ -ray photons of energy greater than 0.7 TeV. So, if such γ -rays were propagating through a dense sea of infrared photons, as implied by previous measurements³⁻⁷, the spectra of the two blazars recorded by HESS would reveal evidence of strong, energy-dependent attenuation. But Aharonian *et al.*¹ show that intergalactic space is more transparent to γ -rays than would be expected if an infrared background excess existed. Remarkably, the attenuation in the HESS images seems consistent merely with the integrated infrared output from resolved foreground galaxies, together with the total extragalactic light produced by second-generation stars

according to recent theoretical calculations¹¹.

What is the cause of the discrepancy between the HESS data and previous results? Aharonian and colleagues point out that their interpretation of the HESS results relies on the assumption that the intrinsic γ -ray spectra of the two active galaxies are not at odds with current models of blazar behaviour. Further high-energy observations of blazars at different cosmological distances should settle this issue. But it seems unlikely that they will be able to supply the fine-tuning required to make the HESS data consistent with the previously reported excess of infrared background light.

The distribution of this excess infrared component over different wavelengths is almost identical to that of sunlight reflected from local interplanetary dust clouds¹². It is therefore conceivable that not all of the foreground emission has been subtracted from the diffuse-sky maps, and the excess may not be extragalactic after all. On the other hand, the fluctuations detected in highly sensitive images taken with the Infrared Array Camera onboard Spitzer⁶ do not change between observations performed six months apart, and changes would be expected if their origin was zodiacal light within the Solar System.

Whatever the final resolution of the mystery concerning the signature of the first stars, the HESS findings¹ will stir much debate among cosmologists. They are likely to spark further attempts to glimpse the crucial early stages of the galaxy formation process.

Piero Madau is in the Department of Astronomy and Astrophysics, University of California, Santa Cruz, California 95064, USA.
e-mail: pmadau@ucolick.org

- Aharonian, F. *et al.* *Nature* **440**, 1018-1021 (2006).
- Spergel, D. N. *et al.* *Astrophys. J.* (submitted); preprint available at www.arxiv.org/astro-ph/0603449 (2006).
- Dwek, E. & Arendt, R. G. *Astrophys. J.* **508**, L9-L12 (1998).
- Gorjian, V., Wright, E. L. & Chary, R. R. *Astrophys. J.* **536**, 550-560 (2000).
- Cambresy, L., Reach, W. T., Beichman, C. A. & Jarrett, T. H. *Astrophys. J.* **555**, 563-571 (2001).
- Kashlinsky, A., Arendt, R. G., Mather, J. & Moseley, S. H. *Nature* **438**, 45-50 (2005).
- Matsumoto, T. *et al.* *Astrophys. J.* **626**, 31-43 (2005).
- Santos, M. R., Bromm, V. & Kamionkowski, M. *Mon. Not. R. Astron. Soc.* **336**, 1082-1092 (2002).
- Madau, P. & Silk, J. *Mon. Not. R. Astron. Soc.* **359**, L37-L41 (2005).
- Salvaterra, R. & Ferrara, A. *Mon. Not. R. Astron. Soc.* **367**, L11-L15 (2006).
- Primack, J. R., Bullock, J. S. & Somerville, R. S. *Am. Inst. Phys. Conf. Proc.* **745**, 23-33 (2005).
- Dwek, E., Arendt, R. G. & Krennrich, F. *Astrophys. J.* **635**, 784-794 (2005).

BRIEF COMMUNICATIONS

A pneumo-hydrostatic skeleton in land crabs

A sophisticated dual support system enables a crab to stay mobile immediately after moulting.

Like their aquatic counterparts, terrestrial crabs repeatedly shed their rigid exoskeleton during moulting. But in the case of land crabs, little water is available to provide a temporary hydrostatic skeleton before the new skeleton hardens, and air does not provide the buoyancy necessary to support the animal. Here we show that whenever its exoskeleton is shed, the blackback land crab *Gecarcinus lateralis* relies on an unconventional type of hydrostatic skeleton that uses both gas and liquid (a 'pneumo-hydrostat'). To our knowledge, this is the first experimental evidence for a locomotor skeleton that depends on a gas. It establishes a new category of hydrostatic skeletal support and possibly a critical adaptation to life on land for the Crustacea.

Arthropods grow by moulting: they secrete a new exoskeleton beneath the old, shed the old skeleton, inflate to a larger size, and wait for the new skeleton to harden^{1–3}. Aquatic crustaceans inflate to a larger size by using water; however, like many insects^{4–7}, the land crab *G. lateralis* inflates its foregut with gas^{8,9}. Recently moulted crabs remain soft for several days before the new skeleton hardens sufficiently to support the forces of muscle contraction. Nevertheless, neither aquatic nor terrestrial crabs are incapacitated during this period.

The aquatic blue crab *Callinectes sapidus* maintains mobility by switching to a hydrostatic skeleton¹⁰ — a fluid-based skeleton that is common in soft-bodied invertebrates¹¹. Hydrostatic skeletons are arranged so that the force of muscle contraction is transmitted by an essentially incompressible aqueous fluid^{11–13}. Muscle contraction increases the pressure in the fluid, causing the deformations or stiffening required for support, movement and locomotion.

We investigated the possibility that the water and air used by *G. lateralis* for inflation might both provide a form of hydrostatic skeletal support, a pneumo-hydrostat, following moulting. First, we simultaneously measured the pressure inside the cheliped (claw) and the force of cheliped flexure. We observed

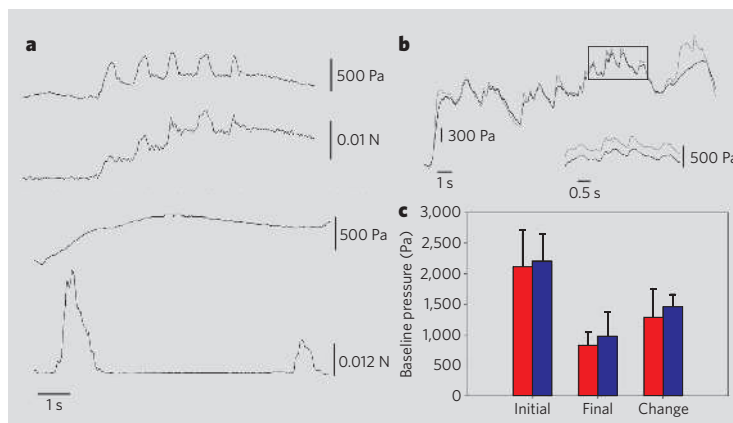


Figure 1 | Haemolymph and gut-gas pressures and movement forces in land crabs after shedding the exoskeleton. **a**, Recordings made 12 h (no exoskeleton; top two traces) and 7 d (with new exoskeleton; bottom two traces) after moulting. Pressure correlates with force in the soft but not in the hard crab. For each pair of traces, the upper shows pressure (Pa, pascals) and the lower shows force (N, newtons). **b**, Pressure inside the cheliped (claw; black trace) and gut (grey trace) of a soft, newly moulted crab. Pressure peaks in the gut and baseline pressure correlate with those in the cheliped. Inset, expanded view of boxed trace. **c**, Average pressure in the cheliped (blue bars) and gut (red bars) before (initial pressure) and after (final pressure) removal of air from the gut. Error bars represent standard error.

a strong correlation between force and pressure in soft, newly moulted crabs but not in hardened crabs, consistent with hydrostatic skeletal support after moulting (Fig. 1a).

We then simultaneously measured the pressure inside the cheliped and gut during cheliped flexure. (For methods, see supplementary information.) In newly moulted crabs, there was a strong correlation between the pressure in the cheliped and gut during cheliped flexure (Fig. 1b). The average baseline pressures were not significantly different (cheliped: 3,792 pascals (Pa), s.d. = 1,029 Pa, $n = 7$; gut: 2,737 Pa, s.d. = 1,329 Pa, $n = 7$; t -test, $P = 0.12$). The average maximum pressures during cheliped flexion were not significantly different either (cheliped: 808 Pa, s.d. = 563 Pa, $n = 14$; gut: 1,088 Pa, s.d. = 510 Pa, $n = 14$; t -test, $P = 0.18$). These results were as expected, because the body is not compartmentalized and therefore local muscle contraction increases the pressure of the haemolymph throughout the body of the crab. As the gut wall is flexible, this results in increased pressure in the gut as well.

We also examined the contribution of gut pressure to the turgidity of the animal. Removing air from the gut causes a decrease in pressure in both the gut and the cheliped (Fig. 1c). The average baseline pressure in the

gut decreased from 2,113 Pa (s.d. = 1,343 Pa, $n = 8$) to 828 Pa (s.d. = 434 Pa, $n = 8$), whereas in the cheliped it decreased from 2,213 Pa (s.d. = 1,318 Pa, $n = 8$) to 981 Pa (s.d. = 736 Pa, $n = 8$). The pressures within the cheliped and gut are not significantly different either before or after the withdrawal of air (initial: $P = 0.89$, s.d. = 1,330; final: $P = 0.64$, s.d. = 604; t -test, $n = 7$). These results support the idea that air in the gut provides skeletal support in crabs, after they have moulted, by increasing their body turgor.

We have shown that a land crab can use a compressible gas in conjunction with an incompressible liquid to provide skeletal support. This gas-liquid skeleton represents a new category of hydrostatic skeleton. The reliance on gas

by a terrestrial arthropod may be more than an adaptation resulting from low water availability: it may also be a biomechanical adaptation to the greater gravitational forces associated with life on land.

Jennifer R. A. Taylor, William M. Kier

Department of Biology, University of North Carolina, Chapel Hill, North Carolina 27599, USA
e-mail: jataylor@email.unc.edu

- Drach, P. *Ann. Inst. Océanogr.* **19**, 103–391 (1939).
- Passano, L. M. in *The Physiology of Crustacea* Vol. 2 (ed. Waterman, T. H.) 473–536 (Academic, New York, 1960).
- Skinner, D. M. in *The Biology of Crustacea* Vol. 9 (eds Bliss, D. E. & Mantel, L. H.) 43–146 (Academic, Orlando, 1985).
- Reynolds, S. E. in *Advances in Insect Physiology* (eds Berridge, M. J., Treherne, J. E. & Wigglesworth, V. B.) 475–595 (Academic, London, 1980).
- Miles, C. I. & Booker, R. J. *Exp. Biol.* **201**, 1785–1798 (1998).
- Hughes, T. D. *Physiol. Entomol.* **5**, 153–164 (1980).
- Cottrell, C. B. *J. Exp. Biol.* **39**, 431–448 (1962).
- Bliss, D. E. *Am. Zool.* **8**, 355–392 (1968).
- Bliss, D. E. *Am. Zool.* **19**, 385–410 (1979).
- Taylor, J. R. A. & Kier, W. M. *Science* **301**, 209–210 (2003).
- Chapman, G. *Biol. Rev.* **33**, 338–371 (1958).
- Chapman, G. *J. Exp. Zool.* **194**, 249–270 (1975).
- Clark, R. B. *Dynamics of Metazoan Evolution* (Clarendon, Oxford, 1964).

Supplementary information accompanies this communication on Nature's website.

Received 17 January; accepted 27 March 2006.

Competing financial interests: declared none.

doi:10.1038/4401005a

Weak pairwise correlations imply strongly correlated network states in a neural population

Elad Schneidman^{1,2,3}, Michael J. Berry II², Ronen Segev² & William Bialek^{1,3}

Biological networks have so many possible states that exhaustive sampling is impossible. Successful analysis thus depends on simplifying hypotheses, but experiments on many systems hint that complicated, higher-order interactions among large groups of elements have an important role. Here we show, in the vertebrate retina, that weak correlations between pairs of neurons coexist with strongly collective behaviour in the responses of ten or more neurons. We find that this collective behaviour is described quantitatively by models that capture the observed pairwise correlations but assume no higher-order interactions. These maximum entropy models are equivalent to Ising models, and predict that larger networks are completely dominated by correlation effects. This suggests that the neural code has associative or error-correcting properties, and we provide preliminary evidence for such behaviour. As a first test for the generality of these ideas, we show that similar results are obtained from networks of cultured cortical neurons.

Much of what we know about biological networks has been learned by studying one element at a time—recording the electrical activity of single neurons, the expression levels of single genes or the concentrations of individual metabolites. On the other hand, important aspects of biological function must be shared among many elements^{1–4}. As a first step beyond the analysis of elements in isolation, much attention has been focused on the pairwise correlation properties of these elements, both in networks of neurons^{5–13} and in networks of genes^{14–16}. But given a characterization of pairwise correlations, what can we really say about the whole network? How can we tell if inferences from a pairwise analysis are correct, or if they are defeated by higher-order interactions among triplets, quadruplets, and larger groups of elements? If these effects are important, how can we escape from the ‘curse of dimensionality’ that arises because there are exponentially many possibilities for such terms?

Here we address these questions in the context of the vertebrate retina, where it is possible to make long, stable recordings from many neurons simultaneously as the system responds to complex, naturalistic inputs^{17–20}. We compare the correlation properties of cell pairs with the collective behaviour in larger groups of cells, and find that the minimal model that incorporates the pairwise correlations provides strikingly accurate but non-trivial predictions of the collective effects. These minimal models are equivalent to the Ising model in statistical physics, and this mapping allows us to explore the properties of larger networks, in particular their capacity for error-correcting representations of incoming sensory data.

The scale of correlations

Throughout the nervous system, individual elements communicate by generating discrete pulses termed action potentials or spikes²¹. If we look through a window of fixed time resolution $\Delta\tau$, then for small $\Delta\tau$ these responses are binary—either the cell spikes (‘1’) or it doesn’t (‘0’). Although some pairs of cells have very strong correlations, most correlations are weak, so that the probability of seeing

synchronous spikes is almost equal to the product of the probabilities of seeing the individual spikes; nonetheless, these weak correlations are statistically significant for most, if not all, pairs of nearby cells. All of these features are illustrated quantitatively by an experiment on the salamander retina (Fig. 1), where we record simultaneously from 40 retinal ganglion cells as they respond to movies taken from a natural setting (see Methods). The correlations between cells have structure on the scale of 20 ms and we use this window as our typical $\Delta\tau$.

The small values of the correlation coefficients suggest an approximation in which the cells are completely independent. For most pairs, this is true with a precision of a few per cent, but if we extrapolate this approximation to the whole population of 40 cells, it fails disastrously. In Fig. 1e, we show the probability $P(K)$ that K of these cells generate a spike in the same small window of duration $\Delta\tau$. If the cells were independent, $P(K)$ would approximate the Poisson distribution, whereas the true distribution is nearly exponential. For example, the probability of $K = 10$ spiking together is $\sim 10^5\times$ larger than expected in the independent model.

The discrepancy between the independent model and the actual data is even more clear if we look at particular patterns of response across the population. Choosing $N = 10$ cells out of the 40, we can form an N -letter binary word to describe the instantaneous state of the network, as in Fig. 1b. The independent model makes simple predictions for the rate at which each such word should occur, and Fig. 1f shows these predictions as a scatter plot against the actual rate at which the words occur in the experiment. At one extreme, the word 1011001010 occurs once per minute, whereas the independent model predicts that this should occur once per three years (a discrepancy of $\sim 10^6\times$). Conversely, the word 1000000010 is predicted to occur once per three seconds, whereas in fact it occurred only three times in the course of an hour. The independent model makes order-of-magnitude errors even for very common patterns of activity, such as a single cell generating a spike while all others are

¹Joseph Henry Laboratories of Physics, ²Department of Molecular Biology, and ³Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08544, USA.

silent. Moreover, within the clusters corresponding to different total numbers of spikes, the predictions and observations are strongly anti-correlated.

We conclude that weak correlations among pairs of neurons coexist with strong correlations in the states of the population as a whole. One possible explanation is that there are specific multi-neuron correlations, whether driven by the stimulus or intrinsic to the network, which simply are not measured by looking at pairs of cells. Searching for such higher-order effects presents many challenges^{22–24}. Another scenario is that small correlations among very many pairs could add up to a strong effect on the network as a whole. If correct, this would be an enormous simplification in our description of the network dynamics.

Minimal consequences of pairwise correlations

To describe the network as a whole, we need to write down a

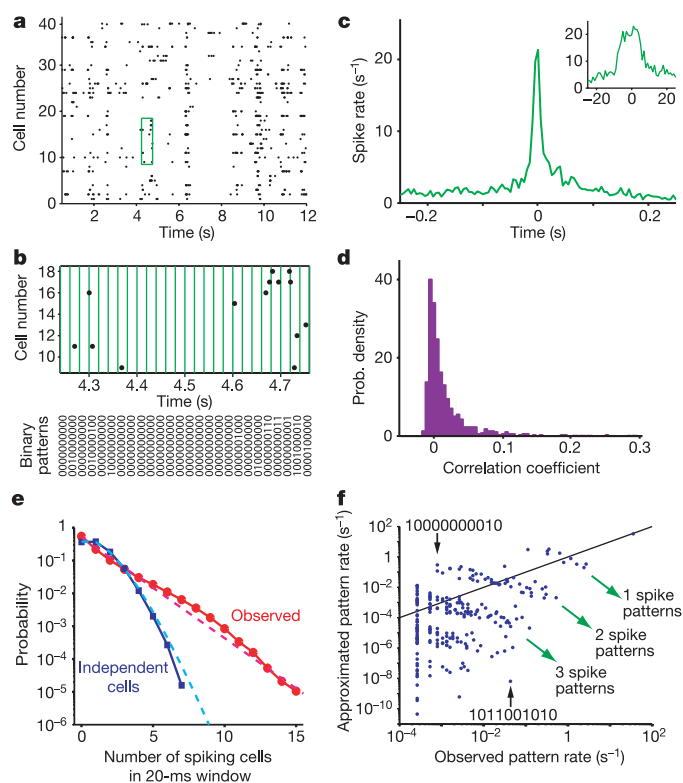


Figure 1 | Weak pairwise cross-correlations and the failure of the independent approximation. **a**, A segment of the simultaneous responses of 40 retinal ganglion cells in the salamander to a natural movie clip. Each dot represents the time of an action potential. **b**, Discretization of population spike trains into a binary pattern is shown for the green boxed area in **a**. Every string (bottom panel) describes the activity pattern of the cells at a given time point. For clarity, 10 out of 40 cells are shown. **c**, Example cross-correlogram between two neurons with strong correlations; the average firing rate of one cell is plotted relative to the time at which the other cell spikes. Inset shows the same cross-correlogram on an expanded time scale; x-axis, time (ms); y-axis, spike rate (s⁻¹). **d**, Histogram of correlation coefficients for all pairs of 40 cells from **a**. **e**, Probability distribution of synchronous spiking events in the 40 cell population in response to a long natural movie (red) approximates an exponential (dashed red). The distribution of synchronous events for the same 40 cells after shuffling each cell's spike train to eliminate all correlations (blue), compared to the Poisson distribution (dashed light blue). **f**, The rate of occurrence of each pattern predicted if all cells are independent is plotted against the measured rate. Each dot stands for one of the $2^{10} = 1,024$ possible binary activity patterns for 10 cells. Black line shows equality. Two examples of extreme mis-estimation of the actual pattern rate by the independent model are highlighted (see the text).

probability distribution for the 2^N binary words corresponding to patterns of spiking and silence in the population. The pairwise correlations tell us something about this distribution, but there are an infinite number of models that are consistent with a given set of pairwise correlations. The difficulty thus is to find a distribution that is consistent only with the measured correlations, and does not implicitly assume the existence of unmeasured higher-order interactions. As the entropy of a distribution measures the randomness or lack of interaction among different variables²⁵, this minimally structured distribution that we are looking for is the maximum entropy distribution²⁶ consistent with the measured properties of individual cells and cell pairs²⁷.

We recall that maximum entropy models have a close connection to statistical mechanics: physical systems in thermal equilibrium are described by the Boltzmann distribution, which has the maximum possible entropy given the mean energy of the system^{26,28}. Thus, any maximum entropy probability distribution defines an energy function for the system we are studying, and we will see that the energy function relevant for our problem is an Ising model. Ising models have been discussed extensively as models for neural networks^{29,30}, but in these discussions the model arose from specific hypotheses

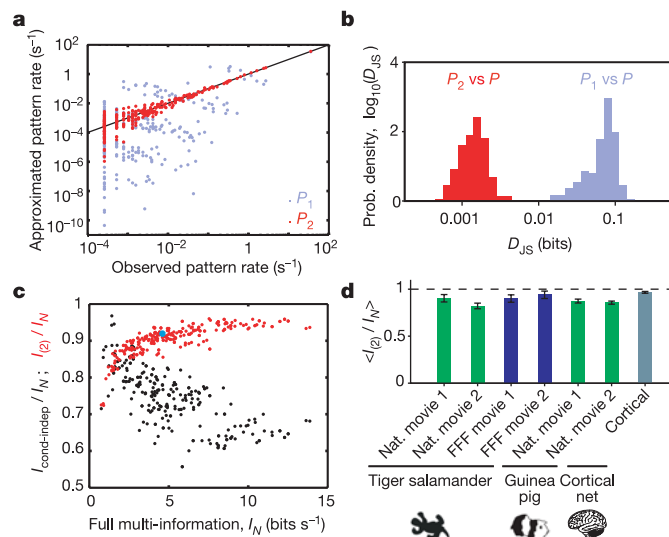


Figure 2 | A maximum entropy model including all pairwise interactions gives an excellent approximation of the full network correlation structure. **a**, Using the same group of 10 cells from Fig. 1, the rate of occurrence of each firing pattern predicted from the maximum entropy model P_2 that takes into account all pairwise correlations is plotted against the measured rate (red dots). The rates of commonly occurring patterns are predicted with better than 10% accuracy, and scatter between predictions and observations is confined largely to rare events for which the measurement of rates is itself uncertain. For comparison, the independent model P_1 is also plotted (from Fig. 1f; grey dots). Black line shows equality. **b**, Histogram of Jensen-Shannon divergences (see Methods) between the actual probability distribution of activity patterns in 10-cell groups and the models P_1 (grey) and P_2 (red); data from 250 groups. **c**, Fraction of full network correlation in 10-cell groups that is captured by the maximum entropy model of second order, $I_{(2)}/I_N$, plotted as a function of the full network correlation, measured by the multi-information I_N (red dots). The multi-information values are multiplied by $1/\Delta\tau$ to give bin-independent units. Every dot stands for one group of 10 cells. The 10-cell group featured in **a** is shown as a light blue dot. For the same sets of 10 cells, the fraction of information of full network correlation that is captured by the conditional independence model, $I_{cond-indep}/I_N$, is shown in black (see the text). **d**, Average values of $I_{(2)}/I_N$ from 250 groups of 10 cells. Results are shown for different movies (see Methods), for different species (see Methods), and for cultured cortical networks; error bars show standard errors of the mean. Similar results are obtained on changing N and $\Delta\tau$; see Supplementary Information.

about the network dynamics. Here, the Ising model is forced upon us as the least-structured model that is consistent with measured spike rates and pairwise correlations; we emphasize that this is not an analogy or a metaphor, but rather an exact mapping.

Whether we view the maximum entropy model through its analogy with statistical physics or simply as a model to be constructed numerically from the data (see Methods), we need meaningful ways of assessing whether this model is correct. Generally, for a network of N elements, we can define maximum entropy distributions P_K that are consistent with all K th-order correlations for any $K = 1, 2, \dots, N$ (ref. 27). These distributions form a hierarchy, from $K = 1$ where all elements are independent, up to $K = N$, which is an exact description that allows arbitrarily complex interactions; their entropies S_K decrease monotonically toward the true entropy S : $S_1 \geq S_2 \geq \dots \geq S_N = S$. The entropy difference or multi-information $I_N = S_1 - S_N$ measures the total amount of correlation in the network, independent of whether it arises from pairwise, triplet or more-complex correlations³¹. The contribution of the K th-order correlation is $I_{(K)} = S_{K-1} - S_K$ and is always positive (more correlation always reduces the entropy); I_N is the sum of all the $I_{(K)}$ (ref. 27). Therefore, the question of whether pairwise correlations provide an effective description of the system becomes the question of whether the reduction in entropy that comes from these correlations, $I_{(2)} = S_1 - S_2$, captures all or most of the multi-information I_N .

Are pairwise correlations enough?

Figure 2 shows the predictions of the maximum entropy model P_2 consistent with pairwise correlations in populations of $N = 10$ cells. Looking in detail at the patterns of spiking and silence of one group of 10 cells, we see that the predicted rates for different binary words

are tightly correlated with the observed rates over a very large dynamic range, so that the dramatic failures of the independent model have been overcome (Fig. 2a).

With 40 cells, we can choose many different populations of 10 cells, and in each case we find that the predicted and observed distributions of words are very similar. It would typically take thousands of independent samples to distinguish reliably between the true distribution of responses and the maximum entropy model, two orders of magnitude more than for the independent model (Fig. 2b).

The success of the pairwise maximum entropy models in capturing the correlation structure of the network is summarized by the fraction $I_{(2)}/I_N \approx 90\%$ (Fig. 2c). This ratio is larger when I_N itself is larger, so that the pairwise model is more effective in describing populations of cells with stronger correlations, and the ability of this model to capture $\sim 90\%$ of the multi-information holds independent of many details (Fig. 2d; see also Supplementary Information): we can vary the particular natural movies shown to the retina, use an artificial movie, change the size of the bins Δt used to define the binary responses, change the number of neurons N that we analyse, and even shift from a lower vertebrate (salamander) to a mammalian (guinea pig) retina. Finally, the correlation structure in a network of cultured cortical neurons³² can be captured by the pairwise model with similar accuracy.

The maximum entropy model describes the correlation structure of the network activity without assumptions about its mechanistic origin. A more traditional approach has been to dissect the correlations into contributions that are intrinsic to the network and those that are driven by the visual stimulus. The simplest model in this view is one in which cells spike independently in response to their input, so that all correlations are generated by covariances of the individual cells' firing rates³³. Although there may be situations in which conditional independence is a good approximation, Fig. 2c shows that this model is less effective than the maximum entropy model in capturing the multi-information for 232 out of 250 groups of 10 neurons, even though the conditional independent model has ~ 200 times more parameters (see Methods). The hypothesis of conditional independence is consistently less effective in capturing the structure of more-strongly correlated groups of cells, which is opposite to the behaviour of the maximum entropy model. Finally, whereas the maximum entropy model can be constructed solely from the observed correlations among neurons, the conditionally independent model requires explicit access to repeated presentations of the visual stimulus. Thus, the central nervous system could learn the maximum entropy model from the data provided by the retina alone, but the conditionally independent model is not biologically realistic in this sense.

We conclude that although the pairwise correlations are small and the multi-neuron deviations from independence are large, the maximum entropy model consistent with the pairwise correlations captures almost all of the structure in the distribution of responses from the full population of neurons. Thus, the weak pairwise correlations imply strongly correlated states. To understand how this happens, it is useful to look at the mathematical structure of the maximum entropy distribution.

Ising models, revisited

We recall that the maximum entropy distribution consistent with a known average energy $\langle E \rangle$ is the Boltzmann distribution, $P \propto \exp(-E/k_B T)$, where k_B is Boltzmann's constant and T is temperature. This generalizes, so that if we know the average values of many variables f_μ describing the system, then the maximum entropy distribution is $P \propto \exp(-\sum_\mu \lambda_\mu f_\mu)$, where there is a separate Lagrange multiplier λ_μ for each constraint^{26,28}. In our case, we are given the average probability of a spike in each cell and the correlations among all pairs. If we represent the activity of cell i by a variable $\sigma_i = \pm 1$, where $+1$ stands for spiking and -1 stands for

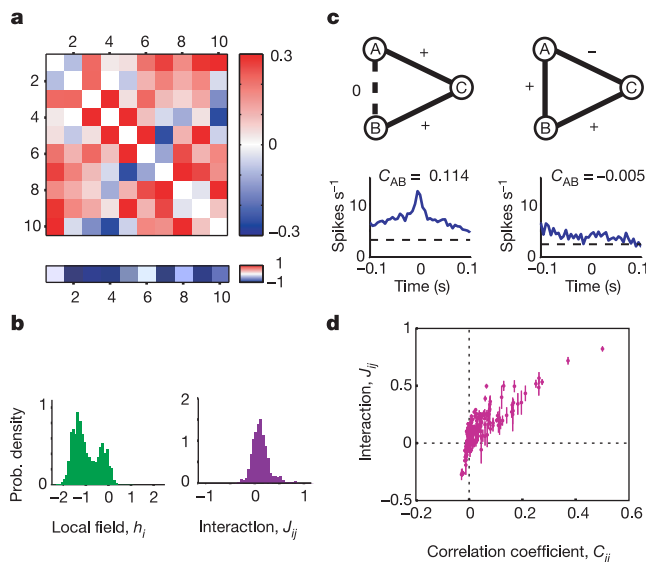


Figure 3 | Pairwise interactions and individual cell biases, as in equation (1). **a**, Example of the pairwise interactions J_{ij} (above) and bias values (or local fields) h_i (below) for one group of 10 cells. **b**, Histograms of h_i and J_{ij} values from 250 different groups of 10 cells. **c**, Two examples of 3 cells within a group of 10. At left, cells A and B have almost no interaction ($J_{AB} = -0.02$), but cell C is very strongly interacting with both A and B ($J_{AC} = 0.52$, $J_{BC} = 0.70$), so that cells A and B exhibit strong correlation, as shown by their cross-correlogram (bottom panel). At right, a 'frustrated' triplet, in which cells A and B have a significant positive interaction ($J_{AB} = 0.13$), as do cells B and C ($J_{BC} = 0.09$), but A and C have a significant negative interaction ($J_{AC} = -0.11$). As a result, there is no clear correlation between cells A and B, as shown by their cross-correlogram (bottom panel). **d**, Interaction strength J_{ij} plotted against the correlation coefficient C_{ij} ; each point shows the value for one cell pair averaged over many different groups of neighbouring cells (190 pairs from 250 groups), and error bars show standard deviations.

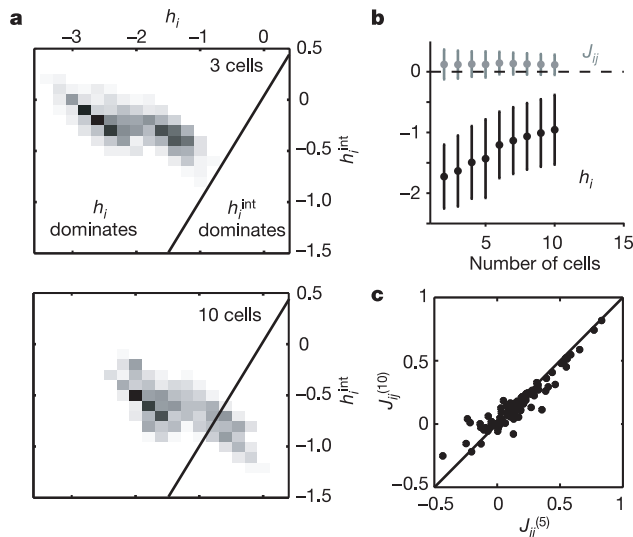


Figure 4 | Interactions and local fields in networks of different size.

a, Greyscale density map of the distribution of effective interaction fields experienced by a single cell h_i^{int} versus its own bias or local field h_i (see the text); distribution formed over network configurations at each point in time during a natural movie for $n = 1,140$ 3-cell groups (top panel) and $n = 250$ 10-cell groups (bottom panel). Black line shows the boundary between dominance of local fields versus interactions. **b**, Mean interactions J_{ij} and local fields h_i , describing groups of N cells, with error bars showing standard deviations across multiple groups. **c**, Pairwise interaction in a network of 10 cells $J_{ij}^{(10)}$ plotted against the interaction values of the same pair in a sub-network containing only 5 cells $J_{ij}^{(5)}$. Line shows equality.

silence, then these constraints are equivalent to fixing the average of each σ_i and the averages of all products $\sigma_i \sigma_j$, respectively. The resulting maximum entropy distribution is

$$P_2(\sigma_1, \sigma_2, \dots, \sigma_N) = \frac{1}{Z} \exp \left[\sum_i h_i \sigma_i + \frac{1}{2} \sum_{i \neq j} J_{ij} \sigma_i \sigma_j \right] \quad (1)$$

where the Lagrange multipliers $\{h_i, J_{ij}\}$ have to be chosen so that the averages $\{\langle \sigma_i \rangle, \langle \sigma_i \sigma_j \rangle\}$ in this distribution agree with experiment; the partition function Z is a normalization factor. This is the Ising model²⁸, where the σ_i are spins, the h_i are local magnetic fields acting on each spin, and the J_{ij} are the exchange interactions; note that $h > 0$ favours spiking and $J > 0$ favours positive correlations.

Figure 3 shows the parameters $\{h_i, J_{ij}\}$ for a particular group of ten cells, as well as the distributions of parameters for many such groups. Most cells have a negative local field, which biases them toward silence. Figure 3c and d illustrates the non-trivial relationship between the pairwise interaction strengths J_{ij} and the observed pairwise correlations.

We can rewrite equation (1) exactly by saying that each neuron or spin σ_i experiences an effective magnetic field that includes the local field or intrinsic bias h_i and a contribution from interactions with all the other spins (neurons), $h_i^{\text{int}} = \frac{1}{2} \sum_{j \neq i} J_{ij} \sigma_j$; note that h_i^{int} depends on whether the other cells are spiking or silent. The intrinsic bias dominates in small groups of cells, but as we look to larger networks, the fact that almost all of the $\sim N^2$ pairs of cells are significantly (if weakly) interacting shifts the balance so that the typical values of the intrinsic bias are reduced while the effective field contributed by the other cells has increased (Fig. 4).

Larger networks and error correction

Groups of $N = 10$ cells are large enough to reveal dramatic departures from independence, but small enough that we can directly sample the relevant probability distributions. What happens at larger N ? In general, we expect that the total capacity of the network to represent its sensory inputs should grow in proportion to the number of neurons, N . This is the usual thermodynamic limit in statistical mechanics, where energy and entropy are proportional to system size²⁸. But this behaviour is not guaranteed when all elements of the system interact with each other. In the Ising model, it is known that if all pairs of spins (here, cells) interact significantly, then to recover the thermodynamic limit the typical size of the interactions J_{ij} must decrease with N (refs 30, 34). Although we have not analysed very large networks, we see no signs of significant changes in J with growing N (Fig. 4b, c).

In a physical system, the maximum entropy distribution is the Boltzmann distribution, and the behaviour of the system depends on the temperature, T . For the network of neurons, there is no real temperature, but the statistical mechanics of the Ising model predicts that when all pairs of elements interact, increasing the number of elements while fixing the typical strength of interactions is equivalent to lowering the temperature, T , in a physical system of fixed size, N . This mapping predicts that correlations will be even more important in larger groups of neurons.

We can see signs of strong correlation emerging by looking at the entropy and multi-information in sub-networks of different sizes. If all cells were independent, the entropy would be S_1 , exactly

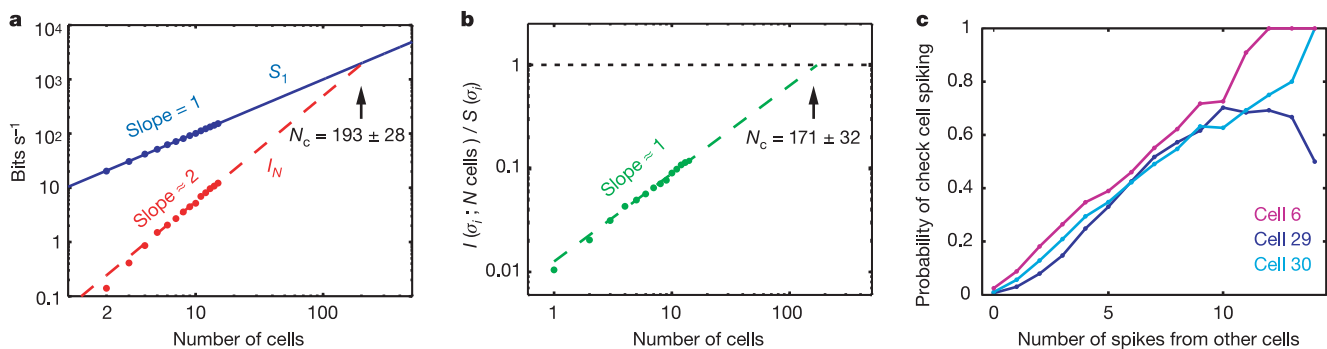


Figure 5 | Extrapolation to larger networks. **a**, Average independent cell entropy S_1 and network multi-information I_N , multiplied by $1/\Delta\tau$ to give bin-independent rates, versus number of cells in the network N . Theoretically, we expect $I_N \propto N(N-1)$ for small N ; the best fit is $I_N \propto N^{1.98 \pm 0.04}$. Extrapolating (dashed line) defines a critical network size N_c , where I_N would be equal to S_1 . **b**, Information that N cells provide about the activity of cell $N+1$, plotted as a fraction of that cell's entropy, $S(\sigma_i)$, versus network size N ; each point is the average value for many different

groups of cells. Extrapolation to larger networks (dashed line, slope = 1.017 ± 0.052) defines another critical network size N_c , where one would get perfect error-correction or prediction of the state of a single cell from the activity of the rest of the network. **c**, Examples of 'check cells', for which the probability of spiking is an almost perfectly linear encoding of the number of spikes generated by the other cells in the network. Cell numbers as in Fig. 1.

proportional to N . For weak correlations, we can solve the Ising model in perturbation theory to show that the multi-information I_N is the sum of mutual information terms between all pairs of cells, and hence $I_N \propto N(N-1)$. This is in agreement with the empirically estimated I_N up to $N=15$, the largest value for which direct sampling of the data provides a good estimate (Fig. 5a), and Monte Carlo simulations of the maximum entropy models suggest that this agreement extends up to the full population of $N=40$ neurons in our experiment (G. Tkačik, E.S., R.S., M.J.B. and W.B., unpublished data). Were this pattern to continue, at $N \approx 200$ cells I_N would become equal to the independent entropy S_1 , and the true entropy $S_N = S_1 - I_N$ would vanish as the system ‘froze’. Because we see variable firing patterns of all the cells, we know that literal freezing of the network into a single state doesn’t happen. On the other hand, networks of $N \approx 200$ cells must be very strongly ordered. Interestingly, experiments indicate that a correlated patch on the retina has roughly this size: the strongest correlations are found for cells within $\sim 200 \mu\text{m}$ of each other, and this area contains ~ 175 ganglion cells in the salamander¹⁹.

Because the interactions J_{ij} have different signs, frustration (Fig. 3c) can prevent the freezing of the system into a single state. Instead there will be multiple states that are local minima of the effective energy function, as in spin glasses³⁴. We find that roughly 40% of all triplets of cells are indeed frustrated. If the number of minimum energy patterns is not too small, then the system retains a significant representational capacity. If the number of patterns is not too large, then observing only some of the cells in the network is sufficient to identify the whole pattern uniquely, just as in the Hopfield model of associative memory²⁹. Thus, the system would have a holographic or error-correcting property, so that an observer who has access only to a fraction of the neurons would nonetheless be able to reconstruct the activity of the whole population.

We can see suggestions of this error-correcting property by asking directly how much the knowledge of activity in N cells tells us about whether cell $N+1$ will spike (Fig. 5b). Our uncertainty about the state of one cell is reduced in proportion to the number of cells that we examine, and if this trend continues, then again at $N \approx 200$ all uncertainty would vanish. Alternatively, we can look for particular kinds of error correction. In our population of 40 cells, we have found three cells for which the probability of spiking is an almost perfectly linear encoding of the number of spikes generated by the other cells in the network (Fig. 5c). To the extent that local field potentials or intrinsic optical signals in cortex reflect the total number of spikes generated by nearby neurons, this observation is analogous to the statement that the spiking in single pyramidal cells is correlated with these more collective responses³⁵. By observing the activity of the ‘check cells’ in Fig. 5c, we can estimate how many spikes are generated by the network as a whole even before we observe any of the other cells’ responses.

Discussion

We have seen that the maximum entropy principle provides a unique candidate model for the whole network that is consistent with observations on pairs of elements but makes no additional assumptions. Despite the opportunity for higher-order interactions in the retina, this model captures more than 90% of the structure in the detailed patterns of spikes and silence in the network, closing the enormous gap between the data and the predictions of a model in which cells fire independently. Because the maximum entropy model has relatively few parameters, we evade the curse of dimensionality and associated sampling problems that would ordinarily limit the exploration of larger networks. The low spiking probabilities, and weak but significant correlations among almost all pairs of cells, are not unique to the retina. Indeed, the maximum entropy model of second order captures over 95% of the multi-information in experiments on cultured networks of cortical neurons. In addition, application of the maximum entropy formalism of ref. 27 to ganglion

cells in monkey retina shows that pairwise correlations in groups of up to $N=8$ ‘ON’ or ‘OFF’ parasol cells, restricted to adjacent cells in each mosaic, can account for 98% of the observed deviations from statistical independence (E. J. Chichilnisky, personal communication).

The success of a model that includes only pairwise interactions provides an enormous simplification in our description of the network. This may be important not only for our analysis, but also for the brain. The dominance of pairwise interactions means that learning rules based on pairwise correlations³⁶ could be sufficient to generate nearly optimal internal models for the distribution of ‘codewords’ in the retinal vocabulary, thus allowing the brain to accurately evaluate new events for their degree of surprise³⁷.

The mapping of the maximum entropy problem to the Ising model, together with the observed level of correlations, implies that groups of $N \approx 200$ cells will behave very differently than smaller groups, and this is especially interesting because the patch of significantly correlated ganglion cells in the retina is close to this critical size¹⁹. Because the response properties of retinal ganglion cells adapt to the input image statistics^{38,39}, this matching of correlation length and correlation strength cannot be an accident of anatomy but rather must be set by adaptive mechanisms. Perhaps there is an optimization principle that determines this operating point, maximizing coding capacity while maintaining the correlation structures that enable error-correction.

Although we have focused on networks of neurons, the same framework has the potential to describe biological networks more generally. In this view, the network is much more than the sum of its parts, but a nearly complete model can be derived from all its pairs.

METHODS

Retinal experiments. Retinae from the larval tiger salamander (*Ambystoma tigrinum*) and the guinea pig (*Cavia porcellus*) were isolated from the eye, retaining the pigment epithelium, and placed over a multi-electrode array¹⁹. Both were perfused with oxygenated medium: room temperature Ringers for salamander and 36 °C Ames medium for guinea pig. Extracellular voltages were recorded by a MultiChannel Systems MEA 60 microelectrode array and streamed to disk for offline analysis. Spike waveforms were sorted either using the spike size and shape from a single electrode¹⁹ or the full waveform on 30 electrodes¹⁸. Recorded ganglion cells were spaced no more than 500 μm apart, and were typically close enough together to have overlapping receptive field centres. Our analysis is based on measurements of 95 cells recorded in 4 salamanders and 35 cells recorded in 2 guinea pigs.

Natural movie clips (‘Nat.’ in Fig. 2) were acquired using a Canon Optura Pi video camera at 30 frames per second. Movie clips broadly sampled woodland scenes as well as man-made environments, and included several qualitatively different kinds of motion: objects moving in a scene, optic flow, and simulated saccades¹⁹. In spatially uniform flicker (‘FFF’ in Fig. 2), the light intensity was randomly chosen to be black or white every 16.7 ms. For most experiments, a 20–30 s stimulus segment was repeated many times; in one experiment, a 16 min movie clip was repeated several times. All visual stimuli were displayed on an NEC FP1370 monitor and projected onto the retina using standard optics. The mean light level was 5 lux, corresponding to photopic vision.

Cultured cortical networks. Data on cultured cortical neurons were recorded by the laboratory of S. Marom (Technion–Israel Institute of Technology) using a multi-electrode array, as described in ref. 32. The data set analysed here is an hour-long epoch of spontaneous neuronal activity recorded through 60 electrodes.

Analysis. Mean spike rates ranged from 0.3 to 4.5 spikes s^{-1} . Spike trains are binned using $\Delta\tau = 20$ -ms windows (unless otherwise noted) into binary sequences of spiking (1) and non-spiking (0); in the rare cases where there is more than one spike in a bin, we denote it as ‘1’. Cross-correlation values were estimated by discretizing the neural response into binary (spike/no spike) variables for each cell, using $\Delta\tau = 20$ -ms bins, and then computing the correlation coefficients among these variables. Because the data sets we consider here are very large (~ 1 hour), the threshold for statistical significance of the individual correlation coefficients is well below $|C| = 0.01$.

Information theoretic quantities such as I_N depend on the full distribution of states for the real system. Estimating these quantities can be difficult, because finite data sets lead to systematic errors⁴⁰. With large data sets (~ 1 hour) and $N < 15$ cells, however, systematic errors are small, and we can use the sample-size dependence of the estimates to correct for these errors, as in ref. 41. For

networks of modest size, as considered here, constructing the maximum entropy distribution consistent with the mean spike rates and pairwise correlations can be viewed as an optimization problem with constraints. Because the entropy is a convex function of the probabilities, and the constraints are linear, many efficient algorithms are available⁴². To test our models we sometimes need surrogate data without correlations. To remove all correlations among neurons (Fig. 1), we shift the whole spike train of each cell by a random time relative to all the other cells. To generate the conditionally independent responses (Fig. 2), we use data from repeated presentations (trials) of the same movie and shuffle the trial labels on each cell independently. We then use the joint probability distribution of the cells under conditional independence, $P_{\text{cond-indep}}(\sigma_1, \sigma_2, \dots, \sigma_N)$, to compute $I_{\text{cond-indep}} = S_1 - S[P_{\text{cond-indep}}(\sigma_1, \sigma_2, \dots, \sigma_N)]$. Note that the conditionally independent model has $NT/\Delta\tau$ parameters, because each cell has its own spike rate, potentially different at each moment in time, where T is the duration of the stimulus movie; in our case $NT/\Delta\tau \approx 10^4$, in contrast to the $N(N+1)/2 = 55$ parameters of the maximum entropy model, for $N = 10$ cells.

The Jensen–Shannon divergence, $D_{\text{JS}}[p||q]$, quantifies the dissimilarity of the distributions p and q , essentially measuring the inverse of the number of independent samples we would need in order to be sure that the two distributions were different⁴³.

Received 22 September 2005; accepted 6 March 2006.

Published online 9 April 2006.

- Hopfield, J. J. & Tank, D. W. Computing with neural circuits: a model. *Science* **233**, 625–633 (1986).
- Georgopoulos, A. P., Schwartz, A. B. & Kettner, R. E. Neuronal population coding of movement direction. *Science* **233**, 1416–1419 (1986).
- Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402** (Suppl. C), 47–52 (1999).
- Barabási, A.-L. & Oltvai, Z. N. Network biology: Understanding the cell's functional organization. *Nature Rev. Genet.* **5**, 101–113 (2004).
- Perkel, D. H. & Bullock, T. H. Neural coding. *Neurosci. Res. Prog. Sum.* **3**, 221–348 (1968).
- Zohary, E., Shadlen, M. N. & Newsome, W. T. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature* **370**, 140–143 (1994).
- Meister, M., Lagnado, L. & Baylor, D. A. Concerted signaling by retinal ganglion cells. *Science* **270**, 1207–1210 (1995).
- Riehle, A., Grun, S., Diesmann, M. & Aertsen, A. Spike synchronization and rate modulation differentially involved in motor cortical function. *Science* **278**, 1950–1953 (1997).
- Dan, Y., Alonso, J. M., Usrey, W. M. & Reid, R. C. Coding of visual information by precisely correlated spikes in the lateral geniculate nucleus. *Nature Neurosci.* **1**, 501–507 (1998).
- Hatsopoulos, N., Ojakangas, C., Paninski, L. & Donoghue, J. Information about movement direction obtained from synchronous activity of motor cortical neurons. *Proc. Natl Acad. Sci. USA* **95**, 15706–15711 (1998).
- Abbott, L. F. & Dayan, P. The effect of correlated variability on the accuracy of a population code. *Neural Comput.* **11**, 91–101 (1999).
- Bair, W., Zohary, E. & Newsome, W. T. Correlated firing in macaque visual area MT: time scales and relationship to behavior. *J. Neurosci.* **21**, 1676–1697 (2001).
- Shamir, M. & Sompolinsky, H. Nonlinear population codes. *Neural Comput.* **16**, 1105–1136 (2004).
- Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
- Alter, O., Brown, P. O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA* **97**, 10101–10106 (2000).
- Holter, N. S., Maritan, A., Cieplak, M., Federoff, N. V. & Banavar, J. R. Dynamic modeling of gene expression data. *Proc. Natl Acad. Sci. USA* **98**, 1693–1698 (2001).
- Meister, M., Pine, J. & Baylor, D. A. Multi-neuronal signals from the retina: acquisition and analysis. *J. Neurosci. Methods* **51**, 95–106 (1994).
- Segev, R., Goodhouse, J., Puchalla, J. L. & Berry, M. J. II. Recoding spikes from a large fraction of the ganglion cells in a retinal patch. *Nature Neurosci.* **7**, 1155–1162 (2004).
- Puchalla, J. L., Schneidman, E., Harris, R. A. & Berry, M. J. II. Redundancy in the population code of the retina. *Neuron* **46**, 492–504 (2005).
- Frechette, E. S. et al. Fidelity of the ensemble code for visual motion in primate retina. *J. Neurophysiol.* **94**, 119–135 (2005).
- Rieke, F., Warland, D., de Ruyter van Steveninck, R. & Bialek, W. *Spikes: Exploring the Neural Code* (MIT Press, Cambridge, 1997).
- Martignon, L. et al. Neural coding: higher-order temporal patterns in the neurostatistics of cell assemblies. *Neural Comput.* **12**, 2621–2653 (2000).
- Grun, S., Diesmann, M. & Aertsen, A. Unitary events in multiple single-neuron spiking activity: I. Detection and significance. *Neural Comput.* **14**, 43–80 (2002).
- Schnitzler, M. J. & Meister, M. Multineuronal firing patterns in the signal from eye to brain. *Neuron* **37**, 499–511 (2003).
- Brillouin, L. *Science and Information Theory* (Academic, New York, 1962).
- Jaynes, E. T. Information theory and statistical mechanics. *Phys. Rev.* **106**, 62–79 (1957).
- Schneidman, E., Still, S., Berry, M. J. II & Bialek, W. Network information and connected correlations. *Phys. Rev. Lett.* **91**, 238701 (2003).
- Landau, L. D. & Lifshitz, E. M. *Statistical Physics* 3rd edn (Pergamon, Oxford, 1980).
- Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl Acad. Sci. USA* **79**, 2554–2558 (1982).
- Amit, D. J. *Modeling Brain Function: The World of Attractor Neural Networks* (Cambridge Univ. Press, Cambridge, UK, 1989).
- Cover, T. M. & Thomas, J. A. *Elements of Information Theory* (Wiley & Sons, New York, 1991).
- Eytan, D., Brenner, N. & Marom, S. Selective adaptation in networks of cortical neurons. *J. Neurosci.* **23**, 9349–9356 (2003).
- Perkel, D. H., Gerstein, G. L. & Moore, G. P. Neuronal spike trains and stochastic point processes. II. Simultaneous spike trains. *Biophys. J.* **7**, 419–440 (1967).
- Mezard, M., Parisi, G. & Virasoro, M. A. *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
- Arieli, A., Sterkin, A., Grinvald, A. & Aertsen, A. Dynamics of ongoing activity: explanation of the large variability in evoked cortical responses. *Science* **273**, 1868–1871 (1996).
- Bi, G. & Poo, M. M. Synaptic modification by correlated activity: Hebb's postulate revisited. *Annu. Rev. Neurosci.* **24**, 139–166 (2001).
- Barlow, H. Conditions for versatile learning, Helmholtz's unconscious inference, and the task of perception. *Vision Res.* **30**, 1561–1571 (1990).
- Smirnakis, S., Berry, M. J. II, Warland, D. K., Bialek, W. & Meister, M. Adaptation of retinal processing to image contrast and spatial scale. *Nature* **386**, 69–73 (1997).
- Hosoya, T., Baccus, S. A. & Meister, M. Dynamic predictive coding by the retina. *Nature* **436**, 71–77 (2005).
- NIPS 2003 Workshop. Estimation of entropy and information of undersampled probability distributions. (<http://nips.cc/Conferences/2003/Workshops/#EstimationofEntropy>) (2003).
- Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. R. & Bialek, W. Entropy and information in neural spike trains. *Phys. Rev. Lett.* **80**, 197–200 (1998).
- Darroch, J. N. & Ratcliff, D. Generalized iterative scaling for log-linear models. *Ann. Math. Stat.* **43**, 1470–1480 (1972).
- Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **37**, 145–151 (1991).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank G. Stephens and G. Tkačik for discussions, N. Tkachuk for help with the experiments, S. Marom for sharing his lab's cultured cortical networks data with us, and E. J. Chichilnisky for sharing his lab's primate retina results with us. This work was supported in part by the NIH and by the E. Matilda Zeigler Foundation.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E.S. (elads@princeton.edu).

Crystal structure of an Hsp90-nucleotide-p23/Sba1 closed chaperone complex

Maruf M. U. Ali¹, S. Mark Roe¹, Cara K. Vaughan¹, Phillipe Meyer^{1†}, Barry Panaretou², Peter W. Piper³, Chrisostomos Prodromou¹ & Laurence H. Pearl¹

Hsp90 (heat shock protein of 90 kDa) is a ubiquitous molecular chaperone responsible for the assembly and regulation of many eukaryotic signalling systems and is an emerging target for rational chemotherapy of many cancers. Although the structures of isolated domains of Hsp90 have been determined, the arrangement and ATP-dependent dynamics of these in the full Hsp90 dimer have been elusive and contentious. Here we present the crystal structure of full-length yeast Hsp90 in complex with an ATP analogue and the co-chaperone p23/Sba1. The structure reveals the complex architecture of the 'closed' state of the Hsp90 chaperone, the extensive interactions between domains and between protein chains, the detailed conformational changes in the amino-terminal domain that accompany ATP binding, and the structural basis for stabilization of the closed state by p23/Sba1. Contrary to expectations, the closed Hsp90 would not enclose its client proteins but provides a bipartite binding surface whose formation and disruption are coupled to the chaperone ATPase cycle.

Hsp90 is an essential molecular chaperone in eukaryotes, required for the activation of many regulatory and signalling 'client' proteins. Hsp90 function depends on its ability to bind and hydrolyse ATP^{1,2}, and pharmacological inhibition by ATP competitors promotes client degradation^{3,4}. The requirement of Hsp90 for the function of oncogenic protein kinases such as ErbB2, Cdk4, B-Raf and Akt/protein kinase B (reviewed in ref. 5) makes it an attractive target for new cancer therapeutics⁶. Hsp90 associates with a plethora of co-chaperones, several of which regulate progress through its ATPase cycle^{7–9}. The p23 co-chaperone and its *Saccharomyces cerevisiae* homologue Sba1 preferentially bind Hsp90 in the presence of ATP or ATP-mimetics^{10–12}. Because p23/Sba1 binds about 70-fold more tightly to ATP-bound Hsp90 (ref. 13), it stabilizes the state required for client-protein activation, slowing ATP turnover^{7,14,15}. This regulatory property explains why, although not essential for Hsp90-dependent activation of client proteins, p23/Sba1 makes it more efficient^{16,17}.

Previous studies suggested that the dimeric Hsp90 operates a 'molecular clamp' mechanism coupled to its ATPase cycle, involving the closure of a 'lid' segment and transient dimerization of the N-terminal nucleotide-binding domain (the N domain) in the ATP-bound state^{18,19}. However, this model has recently been challenged^{20–22}. We have now determined the crystal structure of yeast Hsp90 trapped in a closed conformation, in complex with a non-hydrolysable ATP analogue and p23/Sba1. The structure provides a view of Hsp90 in the ATP-bound state, defining the conformational changes in the N domain that accompany closure and revealing how p23/Sba1 recognizes and stabilizes the ATP-bound conformation of the Hsp90 dimer. The structure confirms the ATPase-coupled molecular clamp mechanism and provides a structural basis from which to understand the ATP-dependent activation of Hsp90 client proteins.

Architecture of the Hsp90-p23/Sba1 complex

Yeast Hsp90, with an Ala107Asn mutation shown to activate the ATPase cycle of Hsp90 (ref. 19) and with truncation of the dispensable charged-linker connecting the N domain and middle segments²³, was crystallized together with the non-hydrolysable ATP analogue AMP-PNP and Sba1, the yeast homologue of p23 (ref. 11). Crystals were phased by molecular replacement with the isolated N-terminal domain²⁴ and middle segment of yeast Hsp90 (ref. 25) and the globular core of human p23 (ref. 26). The structure of the yeast Hsp90 carboxy-terminal domain was determined independently from crystals of a middle-segment and C-terminal Hsp90 construct (M-C) consisting of residues 273–709 refined at 3.0 Å, and completed the molecular replacement solution, allowing refinement of the full approximately 220-kDa (Hsp90)₂-(Sba1)₂ complex against anisotropic data to a maximum resolution of 3.1 Å (3.5 Å isotropic; see Supplementary Information).

The Hsp90-p23/Sba1 complex contains a central Hsp90 dimer with symmetrically disposed p23/Sba1 molecules on each side (Fig. 1a–d). The Hsp90 protomers have a parallel arrangement, with domains arranged in linear order such that the N-terminal domains are at one end of the dimer and the C-terminal domains at the other. The protomers have a left-handed helical twist around the long axis of the dimer. Each Hsp90 protomer consists of four domains (Fig. 1e). The N domain (residues 1–216) is a two-layer sandwich, formed by a twisted β-sheet on one face and a cluster of α-helices on the other, delimiting the binding pocket for ATP and a range of inhibitors^{27–31}. The N domain connects to the middle segment by means of an antiparallel β-strand involving residues 206–213 and 262–269. The loop connecting these strands, formed by the truncated charged-linker, is disordered.

¹Section of Structural Biology, Institute of Cancer Research, Chester Beatty Laboratories, 237 Fulham Road, London SW3 6JB, UK. ²Pharmaceutical Science Research Division, King's College London, Franklin-Wilkins Building, 150 Stamford Street, London SE1 9NN, UK. ³Department of Molecular Biology and Biotechnology, The University of Sheffield, Firth Court, Western Bank, Sheffield S10 2TN, UK. [†]Present address: Laboratoire d'Enzymologie et de Biochimie Structurales, CNRS, Avenue de la Terrasse, 91198 Gif sur Yvette, France.

The middle segment consists of a large three-layer α - β - α domain (residues 273–409) connected to a small α - β - α domain (residues 435–525) by means of a helical coil (residues 411–443)²⁵. The relative orientations of the large and small domains are little changed from the isolated middle-segment structure. An amphipathic loop (residues 329–339), implicated in interactions with client protein²⁵, projects from the inner face of the larger domain towards its equivalent in the other monomer. An extended loop connects the middle segment to the beginning of a curved α -helix (residues 534–559) at the start of the C domain, which consists of a three-stranded β -sheet, packed beneath the curved helix and capping one face of a three-helix coil, whose other face forms the constitutive dimerization interface. A helix-strand segment (residues 587–610) projects from the core of the C domain towards the N-terminal end of the dimer and is involved in dimeric interactions with its equivalent in the other protomer. The yeast C domain is similar in structure to the C domain of *Escherichia coli* HtpG³² but with a significantly different orientation for this projecting helix-strand segment. Residues 678–709, which provide the C-terminal EEVD binding sequence for TPR-domain co-chaperones³³, are disordered.

Sba1 has the same two-layer β -sandwich structure as human p23

(ref. 26), with some differences in the loops connecting the β -strands. Unlike unbound p23, ordered structure is evident beyond the core, with a segment of the charged C-terminal tail (residues 129–139) snaking up towards the middle of the complex.

Domain interactions and conformational changes

Extensive interfaces occur between the domains of the two Hsp90 molecules and the p23/Sba1 molecules. The constitutive dimerization interface between the C domains is essentially identical in the M-C structure (residues 273–709) and the full complex, indicating little change on ATP binding. The interface between the C domain and the small middle-segment domain flexes on going from the unconstrained M-C structure to the ATP-bound conformation, bringing the small middle domains about 10 Å closer together. This displaces the projecting helix-strand segments, which tilt downwards and become less well ordered. As several temperature-sensitive mutations of yeast Hsp90 map in this region^{34,35}, this conformational change may have functional significance. The middle-segment large domains move even more in the ATP-bound conformation, coming about 20 Å closer together. Although the middle segments do not actually come into contact, the gap remaining between them is too small to accommodate a folded client protein (Fig. 2a).

As shown previously^{18,19}, and in contradiction of recent suggestions^{20–22} the N domains form an intimate dimeric interaction,

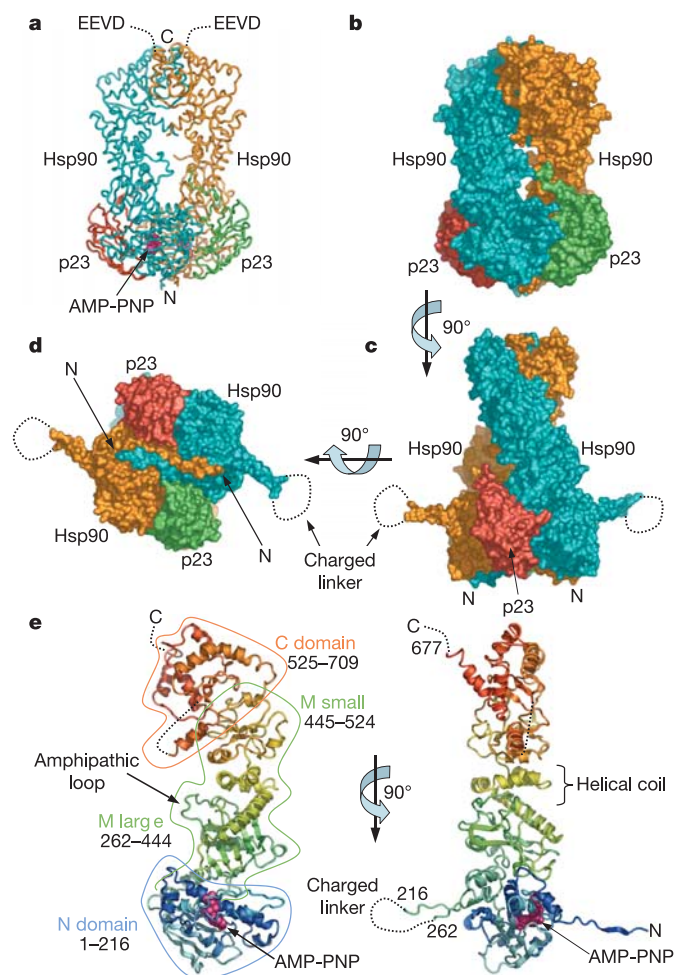


Figure 1 | Architecture of Hsp90-p23/Sba1 complex. **a**, Backbone tracing of the (Hsp90)₂-(p23/Sba1)₂ complex. Blue and orange, Hsp90; red and green, p23/Sba1. All molecular graphics were produced with MacPyMOL (www.pymol.org). **b**, Molecular surface equivalent of **a**. The left-handed twist of the Hsp90 molecules is evident. **c**, As in **b** the projecting β -ribbon is connected by the truncated charged linker, which is disordered in the crystals. **d**, As **c**, but rotated by 90° around the horizontal so that the exchange of the N-terminal strands is evident. **e**, Orthogonal view of Hsp90 protomer, rainbow coloured from the N terminus (blue) to the C terminus (red).

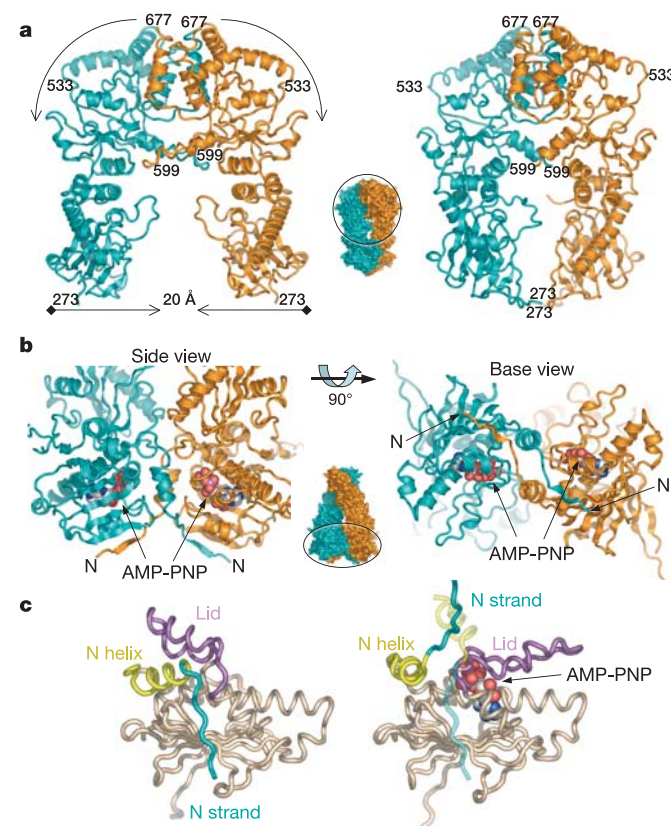


Figure 2 | ATP-dependent conformational changes. **a**, Comparison of the dimerized M-C construct (left), and equivalent regions from the Hsp90-p23/Sba1 complex (right). ATP-dependent association of the N-domains brings the segments about 20 Å closer together. The central 'thumbnail' of the Hsp90 dimer shows the direction of view and location of the M-C region in the overall structure. **b**, Secondary structure diagram of the strand-swapped interface between the N domains of Hsp90, viewed from the base (right) and side (left). **c**, Comparison of the isolated N domain (left) and the AMP-PNP-bound state (right). In addition to lid closure, there is considerable movement of the N-terminal helix and complete detachment of the N-terminal strand, which swaps into the other N domain of the dimer (shown in grey).

involving significant changes in conformation relative to their structure in isolation. First, the N-terminal β -strand (residues 1–9) swaps over to hydrogen-bond to the edge of the main β -sheet in the N domain of the other monomer, with concomitant movement of the first α -helix (residues 13–22)—an arrangement also seen in dimers of MutL and GyrB^{36,37} (Fig. 2b). Second, the ‘lid’ segment (residues 94–125) swings through nearly 180° from its ‘open’ position in the isolated N-domain, hinged at Gly 94 and Gly 121, to fold over the nucleotide-binding pocket. The lid movement exposes a hydrophobic patch centred on Leu 15 and Leu 18, which along with Gln 14, and Thr 95, Ile 96, Ala 97 and Phe 120 from the hinge-points of the lid, form a substantial interface with their equivalents in the other monomer. Thus, as we previously suggested, lid closure in the ATP-bound state reveals a hydrophobic patch whose burial stabilizes N-domain association¹⁹. The lid movement explains the effects of the Thr101Ile and Ala107Asn mutations. Thr 101 is buried in the open conformation of the lid, becoming exposed in the ATP-bound complex; its mutation to the more hydrophobic isoleucine stabilizes the open conformation, decreasing ATPase activity. Ala 107 comes close to Tyr 47 in the closed conformation; mutation to asparagine allows a polar interaction with Tyr 47, stabilizing the closed conformation and enhancing ATPase activity (Fig. 2c).

Although the middle segments move together, they do not make contact. However, each interacts with the N domain of the other protomer, in addition to the interface with its own N domain (see below). This intermolecular inter-domain interface involves Thr 22, Val 23 and Tyr 24 from the N domain, in a hydrophobic interaction with Leu 376 and Leu 378 in the middle-segment loop bearing the catalytic residue Arg 380. The involvement of Thr 22 in this interface is particularly interesting, because Thr22Ile had been isolated as a *ts* allele *in vivo*³⁵ and activated the ATPase of Hsp90 *in vitro*¹⁹, although the mechanism of this was not clear. Mutation to the more hydrophobic isoleucine would reinforce the interface with the catalytic loop, stabilizing its active conformation (Fig. 3a).

Nucleotide interactions and active-site formation

The adenine, ribose, α -phosphate and β -phosphate groups of AMP-PNP occupy the same position as ADP in high-resolution studies of the isolated N-terminal domain^{2,27}. The γ -phosphate is cradled in a glycine-rich loop at the C-terminal hinge of the lid, hydrogen bonding with the peptide backbone of Gln 119, Gly 121, Val 122 and Gly 123. The β -phosphate hydrogen bonds with the peptides of Phe 124, and Gly 100 from the N-terminal hinge (Fig. 3b).

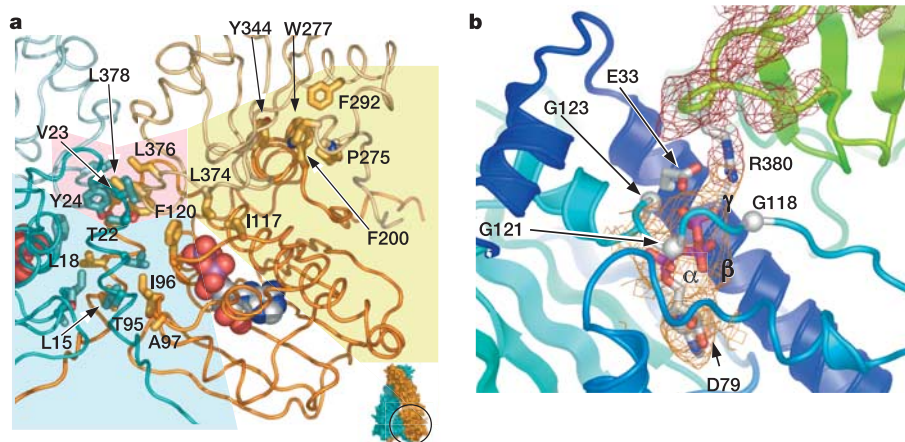


Figure 3 | Domain interfaces and active-site formation. **a**, Interfaces between the Hsp90 N domains and middle segments: the intermolecular N domain interface (pale blue background), the intermolecular N domain and middle segment interface (pale pink), and the intramolecular N domain and middle segment interface (pale yellow) that assembles the catalytic apparatus. **b**, Details of the split catalytic apparatus. The γ -phosphate of

The nucleotide makes a single contact outside the N domain—a polar interaction between its γ -phosphate and the head group of Arg 380, which projects into the nucleotide-binding pocket from the middle segment, close to the catalytically essential Glu 33 (ref. 1). The N domain makes extensive interactions with the middle segment of the same protomer, involving the side chain of Phe 200 from the N domain, packing into a hydrophobic pocket formed by Thr 273, Pro 275, Trp 277, Phe 292 and Tyr 344 in the middle segment. Residues 114–120 from the lid in the N domain make polar and hydrophobic interactions with residues 372–379 from the middle-segment catalytic loop. Both these segments have different conformations in the isolated domains, so their conformations in the ATP-bound complex are mutually interdependent (Fig. 3a).

Previous studies suggested close interactions between the N domain and middle segment of the protein in the ATP-bound state, and identified Arg 380 as an essential catalytic residue^{25,38}. The juxtaposition of the N domain and middle segments, and the proximity of Arg 380 to the bound nucleotide in the structure presented here, completely vindicate that mechanistic proposal. However, the involvement of Arg 380 has recently been questioned on the basis of structures of the *E. coli* Hsp90 HtpG, lacking the C-terminal dimerization domain (residues 1–559), and in the presence of ADP²⁰. These structures show radically different conformations from those described here, with little contact between the middle segment and the nucleotide-binding pocket, and with the equivalent of Arg 380 too far from the nucleotide to have a catalytic function. Although the bacterial homologue might operate by a different mechanism, the strong conservation of the residues involved in the interface between the N domain and the middle segment, and the sequence identity of almost all residues involved in nucleotide interactions, make this highly improbable. Most probably the described conformation of HtpG is an artefact, stabilized by favourable crystal lattice contacts and reflecting the considerable flexibility of the protein in the ATP-free state.

Structural basis for conformation-dependent recruitment of p23/Sba1

Each p23/Sba1 molecule lies in a depression at the junction of the two N domains of Hsp90, contacting each by means of distinct surface patches (Fig. 4a). The smaller interface involves residues 31–37 and 85–91 at the tips of β -hairpins in p23/Sba1 and residues 12–21 and 151–155 from one Hsp90 N domain. The larger interface involves residues 13–16 from the N-terminal strand of p23/Sba1 along with a

AMP-PNP is orientated for attack by a water activated by Glu 33. Arg 380 polarizes the γ - β phosphodiester bond and neutralizes the transition state. Mutation of Glu 33 or Arg 380 kills catalysis; mutation of Asp 79 abolishes ATP binding^{1,25}. The mesh shows $F_o - F_c$ 'omit' electron density for the catalytic loop (red) and the ATP analogue (orange).

broad loop formed by residues 113–118, which packs against the surface of the lid segment of the other N domain (Fig. 4b). Formation of this interface requires lid closure, so that the binding of p23/Sba1 would stabilize the ATP-bound conformation. Consistent with this is the observation that mutation of Ile 117 in this loop causes significant loss of Sba1 function *in vivo*³⁹.

A third interface involves residues 120–125 of p23/Sba1 binding across the intermolecular inter-domain junction (see above). One side of this bridging interaction involves Lys 27 and Lys 387 of Hsp90, and Asp 122 and the peptide carbonyls of Phe 121, Asp 122 and Trp 124, of p23/Sba1. On the other side, Phe 121 and Trp 124 from p23/Sba1 pack into a hydrophobic recess formed by Leu 315, Pro 375, Ile 388 and Val 391 from Hsp90 (Fig. 4c). Lys 387, Ile 388 and Val 391 from the catalytic loop make very similar interactions with Aha1 (ref. 38), an Hsp90 co-chaperone that significantly activates the inherent ATPase of Hsp90^{7,13,40} (Fig. 4d). Binding of Aha1 to Hsp90 remodels the catalytic loop towards the conformation seen in the AMP-PNP-bound Hsp90–p23/Sba1 complex. Although unrelated, Aha1 and p23/Sba1 share an ability to stabilize the catalytic loop in an active conformation, promoting ATP hydrolysis. However, with a common binding site, their interaction with Hsp90 is mutually exclusive⁴¹.

None of the Hsp90–p23/Sba1 interfaces have significant affinity individually, but in combination they add to a substantial degree of interaction. However, the p23/Sba1-binding sites are presented in the correct three-dimensional configuration only when Hsp90 is in the ATP-bound state. Thus, the recruitment of p23/Sba1 is ATP dependent, but once bound it provides a scaffold, reinforcing the ATP-bound conformation and extending the lifetime of the state of the chaperone required for client-protein activation.

ATPase coupled conformational cycle and client-protein activation

The structure presented here defines the ‘tense’ state of the chaperone and confirms the proposed dimerization-coupled ‘split’ ATPase

mechanism^{18,19}, uniting Hsp90 with other dimeric GHKL ATPases such as MutL and GyrB^{36,37}. Because the key residues involved are highly conserved between yeast and human Hsp90s, it is difficult to rationalize suggestions that human Hsp90 has a different mechanism²². Hsp90 in solution does not have a single ‘relaxed’ conformation but exists as a continuum of C-terminally dimerized conformations in which the rest of the molecule is unconstrained⁴². By contrast, the ATP-bound state is a highly constrained structure whose formation involves coupled conformational switches in the N terminus and lid of the N domain, and in the catalytic loop of the middle segment, which together position the two halves of the catalytic apparatus for ATP hydrolysis. These switches, rather than ATP hydrolysis itself, govern the rate of the chaperone cycle; mutations that affect them alter Hsp90 ATPase activity *in vitro* and client-protein activation *in vivo*^{19,34,35,43}. Furthermore, co-chaperones of Hsp90 such as p23/Sba1, Aha1 and Cdc37, which regulate the chaperone ATPase cycle, do so by interacting with these switch components^{13,38,44}.

Mutational analysis^{1,2}, and the effect of ATP-competitive inhibitors⁶, show that the binding of ATP to Hsp90 is crucial for client-protein activation *in vivo*. Now the structural consequences of ATP binding to Hsp90 are understood, the key question remains: What property of this conformation contributes to client-protein activation? At least for protein kinases some insight comes from an electron microscopic reconstruction of an Hsp90–Cdc37–Cdk4 complex (C. K. Vaughan, U. Gohlke, F. Sobott, V. M. Good, M. M. U. Ali, C. Prodromou, C. V. Robinson, H. R. Saibil and L. H. Pearl, manuscript submitted). This suggests that the two lobes of the kinase interact with different domains of Hsp90, so that the client conformation is directly coupled to the changes in their relative position that accompany the chaperone ATPase cycle. Although this may provide a general mechanism for coupling chaperone and client conformations, how such changes in a client facilitate its activation remains to be defined. The structural data presented here provide a firm basis from which this key mechanistic question can be addressed.

METHODS

A full-length construct of yeast Hsp90 harbouring the mutation Ala107Asn and with residues 221–255 in the charged linker region deleted and replaced by LQHMASVD; an N-terminal and middle-segment (M-C) construct of yeast Hsp90 (residues 273–709); and full-length yeast p23 (Sba1) were inserted with the addition of an N-terminal His₆ tag and PreScission protease cleavage site into pRSETA and expressed in *E. coli* BL21 (DE3) pLysS. The expressed proteins were purified to homogeneity by column chromatography and used in crystallization experiments (see Supplementary Methods).

Crystals of the Hsp90–p23/Sba1–AMP-PNP complex and of the M-C construct were grown in hanging-drop vapour diffusion experiments, snap-frozen in liquid nitrogen, and used to collect X-ray diffraction data at the European Synchrotron Radiation Facility, Grenoble. The M-C structure was solved by molecular replacement with the yeast middle-segment structure, and the C-terminal domain was built from difference Fourier maps. The complex structure was determined by molecular replacement with structures of the separate N, middle and C domains of yeast Hsp90, and with human p23. Structures were refined with automated procedures interspersed with manual rebuilding (see Supplementary Methods).

Received 20 January; accepted 9 March 2006.

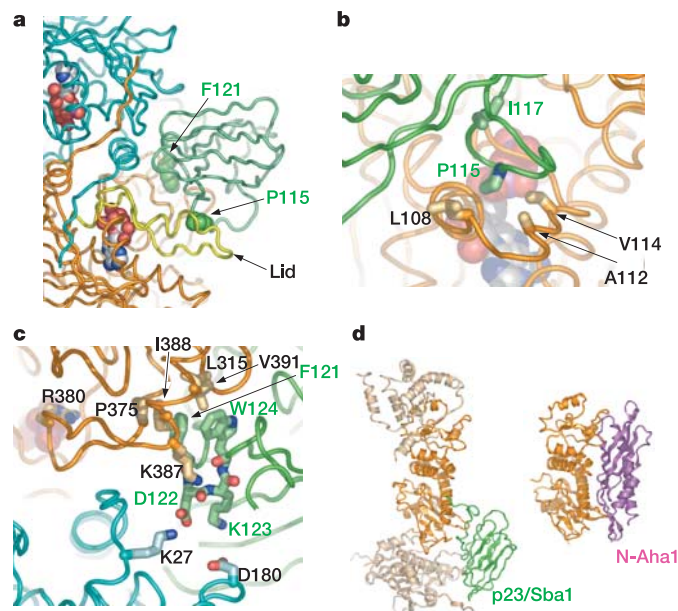


Figure 4 | Mechanism of ATPase regulation by p23/Sba1. **a**, An overall view is shown of one p23/Sba1 molecule (green) packed against the Hsp90 dimer (cyan and orange), and interacting with the ‘lid’ (yellow) in its ATP-dependent closed conformation. **b**, **c**, Details of the general view shown in **a**. **b**, Residues 113–118 of p23/Sba1 (green) pack onto the ‘lid’ (orange). Mutation of Ile 117 causes severe loss of function of Sba1 in yeast³⁹. **c**, The conserved DFxxW motif of p23/Sba1 binds across the intermolecular inter-domain interface, reinforcing the activated conformation of the catalytic loop carrying Arg 380. **d**, Binding sites for p23/Sba1 (left) and Aha1 (right) overlap, so their recruitment to Hsp90 is mutually exclusive⁴¹.

1. Panaretou, B. *et al.* ATP binding and hydrolysis are essential to the function of the Hsp90 molecular chaperone *in vivo*. *EMBO J.* **17**, 4829–4836 (1998).
2. Obermann, W. M. J., Sonderrmann, H., Russo, A. A., Pavletich, N. P. & Hartl, F. U. *In vivo* function of Hsp90 is dependent on ATP binding and ATP hydrolysis. *J. Cell Biol.* **143**, 901–910 (1998).
3. Mimnaugh, E. G., Chavany, C. & Neckers, L. Polyubiquitination and proteasomal degradation of the p185c-erbB-2 receptor protein-tyrosine kinase induced by geldanamycin. *J. Biol. Chem.* **271**, 22796–22801 (1996).
4. Schneider, C. *et al.* Pharmacologic shifting of a balance between protein folding and degradation mediated by Hsp90. *Proc. Natl Acad. Sci. USA* **93**, 14536–14541 (1996).
5. Pearl, L. H. Hsp90 and Cdc37—a chaperone cancer conspiracy. *Curr. Opin. Genet. Dev.* **15**, 55–61 (2005).

6. Workman, P. Combinatorial attack on multistep oncogenesis by inhibiting the Hsp90 molecular chaperone. *Cancer Lett.* **206**, 149–157 (2004).
7. Panaretou, B. *et al.* Activation of the ATPase activity of Hsp90 by the stress-regulated co-chaperone Aha1. *Mol. Cell* **10**, 1307–1318 (2002).
8. Prodromou, C. *et al.* Regulation of Hsp90 ATPase activity by tetratricopeptide repeat (TPR)-domain co-chaperones. *EMBO J.* **18**, 754–762 (1999).
9. Siligardi, G. *et al.* Regulation of Hsp90 ATPase activity by the co-chaperone Cdc37p/p50^{cdc37}. *J. Biol. Chem.* **277**, 20151–20159 (2002).
10. Johnson, J. L. & Toft, D. O. Binding of p23 and hsp90 during assembly with the progesterone receptor. *Mol. Endocrinol.* **9**, 670–678 (1995).
11. Fang, Y., Fliss, A. E., Rao, J. & Caplan, A. J. SBA1 encodes a yeast hsp90 cochaperone that is homologous to vertebrate p23 proteins. *Mol. Cell. Biol.* **18**, 3727–3734 (1998).
12. Sullivan, W. *et al.* Nucleotides and two functional states of hsp90. *J. Biol. Chem.* **272**, 8007–8012 (1997).
13. Siligardi, G. *et al.* Co-chaperone regulation of conformational switching in the Hsp90 ATPase cycle. *J. Biol. Chem.* **279**, 51989–51998 (2004).
14. McLaughlin, S. H., Smith, H. W. & Jackson, S. E. Stimulation of the weak ATPase activity of human Hsp90 by a client protein. *J. Mol. Biol.* **315**, 787–798 (2002).
15. Richter, K., Walter, S. & Buchner, J. The co-chaperone Sba1 connects the ATPase reaction of Hsp90 to the progression of the chaperone cycle. *J. Mol. Biol.* **342**, 1403–1413 (2004).
16. Pratt, W. B. & Dittmar, K. D. Studies with purified chaperones advance the understanding of the mechanism of glucocorticoid receptor hsp90 heterocomplex assembly. *Trends Endocrinol. Metab.* **9**, 244–252 (1998).
17. Young, J. C. & Hartl, F. U. Polypeptide release by Hsp90 involves ATP hydrolysis and is enhanced by the co-chaperone p23. *EMBO J.* **19**, 5930–5940 (2000).
18. Chadli, A. *et al.* Dimerization and N-terminal domain proximity underlie the function of the molecular chaperone heat shock protein 90. *Proc. Natl Acad. Sci. USA* **97**, 12524–12529 (2000).
19. Prodromou, C. *et al.* The ATPase cycle of Hsp90 drives a molecular ‘clamp’ via transient dimerization of the N-terminal domains. *EMBO J.* **19**, 4383–4392 (2000).
20. Huai, Q. *et al.* Structures of the N-terminal and middle domains of *E. coli* Hsp90 and conformation changes upon ADP binding. *Structure* **13**, 579–590 (2005).
21. Immormino, R. M. *et al.* Ligand-induced conformational shift in the N-terminal domain of GRP94, an Hsp90 chaperone. *J. Biol. Chem.* **279**, 46162–46171 (2004).
22. McLaughlin, S. H., Ventouras, L. A., Lobbezoo, B. & Jackson, S. E. Independent ATPase activity of Hsp90 subunits creates a flexible assembly platform. *J. Mol. Biol.* **344**, 813–826 (2004).
23. Louvion, J. F., Warth, R. & Picard, D. Two eukaryote-specific regions of Hsp82 are dispensable for its viability and signal transduction functions in yeast. *Proc. Natl Acad. Sci. USA* **93**, 13937–13942 (1996).
24. Prodromou, C., Roe, S. M., Piper, P. W. & Pearl, L. H. A molecular clamp in the crystal structure of the N-terminal domain of the yeast Hsp90 chaperone. *Nature Struct. Biol.* **4**, 477–482 (1997).
25. Meyer, P. *et al.* Structural and functional analysis of the middle segment of Hsp90: Implications for ATP hydrolysis and client-protein and co-chaperone interactions. *Mol. Cell* **11**, 647–658 (2003).
26. Weaver, A. J., Sullivan, W. P., Felts, S. J., Owen, B. A. & Toft, D. O. Crystal structure and activity of human p23, a heat shock protein 90 co-chaperone. *J. Biol. Chem.* **275**, 23045–23052 (2000).
27. Prodromou, C. *et al.* Identification and structural characterization of the ATP/ADP-binding site in the Hsp90 molecular chaperone. *Cell* **90**, 65–75 (1997).
28. Stebbins, C. E. *et al.* Crystal structure of an Hsp90-geldanamycin complex: Targeting of a protein chaperone by an antitumor agent. *Cell* **89**, 239–250 (1997).
29. Roe, S. M. *et al.* The structural basis for inhibition of the Hsp90 molecular chaperone, by the anti-tumour antibiotics radicicol and geldanamycin. *J. Med. Chem.* **42**, 260–266 (1999).
30. Cheung, K. M. *et al.* The identification, synthesis, protein crystal structure and *in vitro* biochemical evaluation of a new 3,4-diarylpyrazole class of Hsp90 inhibitors. *Bioorg. Med. Chem. Lett.* **15**, 3338–3343 (2005).
31. Wright, L. *et al.* Structure-activity relationships in purine-based inhibitor binding to HSP90 isoforms. *Chem. Biol.* **11**, 775–785 (2004).
32. Harris, S. F., Shiau, A. K. & Agard, D. A. The crystal structure of the carboxy-terminal dimerization domain of htpG, the *Escherichia coli* Hsp90, reveals a potential substrate binding site. *Structure* **12**, 1087–1097 (2004).
33. Scheufler, C. *et al.* Structure of TPR domain-peptide complexes: critical elements in the assembly of the Hsp70-Hsp90 multichaperone machine. *Cell* **101**, 199–210 (2000).
34. Bohren, S. P. & Yamamoto, K. R. Isolation of Hsp90 mutants by screening for decreased steroid receptor function. *Proc. Natl Acad. Sci. USA* **90**, 11424–11428 (1993).
35. Nathan, D. F. & Lindquist, S. Mutational analysis of Hsp90 function: interactions with a steroid receptor and a protein kinase. *Mol. Cell. Biol.* **15**, 3917–3925 (1995).
36. Ban, C., Junop, M. & Yang, W. Transformation of MutL by ATP binding and hydrolysis: a switch in DNA mismatch repair. *Cell* **97**, 85–97 (1999).
37. Wigley, D. B., Davies, G. J., Dodson, E. J., Maxwell, A. & Dodson, G. Crystal structure of an N-terminal fragment of the DNA gyrase B protein. *Nature* **351**, 624–629 (1991).
38. Meyer, P. *et al.* Structural basis for recruitment of the ATPase activator Aha1 to the Hsp90 chaperone machinery. *EMBO J.* **23**, 511–519 (2004).
39. Oxelmark, E. *et al.* Genetic dissection of p23, an Hsp90 cochaperone, reveals a distinct surface involved in estrogen receptor signaling. *J. Biol. Chem.* **278**, 36547–36555 (2003).
40. Lotz, G. P., Lin, H., Harst, A. & Obermann, W. M. J. Aha1 binds to the middle domain of Hsp90, contributes to client protein activation, and stimulates the ATPase activity of the molecular chaperone. *J. Biol. Chem.* **278**, 17228–17235 (2003).
41. Harst, A., Lin, H. & Obermann, W. M. J. Aha1 competes with Hop, p50 and p23 for binding to the molecular chaperone Hsp90 and contributes to kinase and hormone receptor activation. *Biochem. J.* **387**, 789–796 (2005).
42. Zhang, W. *et al.* Biochemical and structural studies of the interaction of Cdc37 with Hsp90. *J. Mol. Biol.* **340**, 891–907 (2004).
43. Richter, K., Reinstein, J. & Buchner, J. N-terminal residues regulate the catalytic efficiency of the Hsp90 ATPase cycle. *J. Biol. Chem.* **277**, 44905–44910 (2002).
44. Roe, S. M. *et al.* The mechanism of Hsp90 regulation by the protein kinase-specific cochaperone p50^{cdc37}. *Cell* **116**, 87–98 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Gold, P. Wan, A. Doré and M. Kilkenny for assistance with synchrotron data collection. This work was supported by a Programme Grant from The Wellcome Trust and infrastructural support for structural biology at The Institute of Cancer Research, from Cancer Research UK.

Author Information Coordinates and structure factors for the Hsp90-p23/Sba1-AMP-PNP complex and M-C construct have been deposited in the Protein Data Bank with accession codes 2CGE and 2CG9. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to L.H.P. (laurence.pearl@icr.ac.uk).

LETTERS

A low level of extragalactic background light as revealed by γ -rays from blazars

F. Aharonian¹, A. G. Akhperjanian², A. R. Bazer-Bachi³, M. Beilicke⁴, W. Benbow¹, D. Berge¹, K. Bernlöhr^{1,5}, C. Boisson⁶, O. Bolz¹, V. Borrel³, I. Braun¹, F. Breitling⁵, A. M. Brown⁷, P. M. Chadwick⁷, L.-M. Chouinet⁸, R. Cornils⁴, L. Costamante^{1,9}, B. Degrange⁸, H. J. Dickinson⁷, A. Djannati-Atai¹⁰, L. O'C. Drury¹¹, G. Dubus⁸, D. Emmanoulopoulos¹², P. Espigat¹⁰, F. Feinstein¹³, G. Fontaine⁸, Y. Fuchs¹⁴, S. Funk¹, Y. A. Gallant¹³, B. Giebels⁸, S. Gillessen¹, J. F. Glicenstein¹⁵, P. Goret¹⁵, C. Hadjichristidis⁷, D. Hauser¹, M. Hauser¹², G. Heinzlmann⁴, G. Henri¹⁴, G. Hermann¹, J. A. Hinton¹, W. Hofmann¹, M. Holleran¹⁶, D. Horns¹, A. Jacholkowska¹³, O. C. de Jager¹⁶, B. Khélifi¹, S. Klages¹, Nu. Komin⁵, A. Konopelko⁵, I. J. Latham⁷, R. Le Gallou⁷, A. Lemièrre¹⁰, M. Lemoine-Goumard⁸, N. Leroy⁸, T. Lohse⁵, J. M. Martin⁶, O. Martineau-Huynh¹⁷, A. Marcowith³, C. Masterson^{1,9}, T. J. L. McComb⁷, M. de Naurois¹⁷, S. J. Nolan⁷, A. Noutsos⁷, K. J. Orford⁷, J. L. Osborne⁷, M. Ouchrif^{9,17}, M. Panter¹, G. Pelletier¹⁴, S. Pita¹⁰, G. Pühlhofer¹², M. Punch¹⁰, B. C. Raubenheimer¹⁶, M. Raue⁴, J. Raux¹⁷, S. M. Rayner⁷, A. Reimer¹⁸, O. Reimer¹⁸, J. Ripken⁴, L. Rob¹⁹, L. Rolland¹⁷, G. Rowell¹, V. Sahakian², L. Saugé¹⁴, S. Schlenker⁵, R. Schlickeiser¹⁸, C. Schuster¹⁸, U. Schwanke⁵, M. Siewert¹⁸, H. Sol⁶, D. Spangler⁷, R. Steenkamp²⁰, C. Stegmann⁵, J.-P. Tavernet¹⁷, R. Terrier¹⁰, C. G. Théoret¹⁰, M. Tluczykont^{8,9}, C. van Eldik¹, G. Vasileiadis¹³, C. Venter¹⁶, P. Vincent¹⁶, H. J. Völk¹ & S. J. Wagner¹²

The diffuse extragalactic background light consists of the sum of the starlight emitted by galaxies through the history of the Universe, and it could also have an important contribution from the 'first stars', which may have formed before galaxy formation began. Direct measurements are difficult and not yet conclusive, owing to the large uncertainties caused by the bright foreground emission associated with zodiacal light¹. An alternative approach^{2–5} is to study the absorption features imprinted on the γ -ray spectra of distant extragalactic objects by interactions of those photons with the background light photons⁶. Here we report the discovery of γ -ray emission from the blazars⁷ H 2356–309 and 1ES 1101–232, at redshifts $z = 0.165$ and $z = 0.186$, respectively. Their unexpectedly hard spectra provide an upper limit on the background light at optical/near-infrared wavelengths that appears to be very close to the lower limit given by the integrated light of resolved galaxies⁸. The background flux at these wavelengths accordingly seems to be strongly dominated by the direct starlight from galaxies, thus excluding a large contribution from other sources—in particular from the first stars formed⁹. This result also indicates that intergalactic space is more transparent to γ -rays than previously thought.

The observations were carried out with the High Energy Stereoscopic System¹⁰ (HESS), a system of four imaging atmospheric Cherenkov telescopes operating at energies $E \geq 0.1$ TeV. These two blazars are at present the most distant sources for which spectra have

been measured at these energies (Table 1).

Intergalactic absorption is caused by the process of photon-photon collision and pair production. The original spectrum emitted by the source (which we call 'intrinsic') is modified such that the observed flux $F_{\text{obs}}(E) = F_{\text{int}}(E) \times e^{-\tau(E)}$, where the optical depth $\tau(E)$ depends on the spectral energy distribution (SED) of the extragalactic background light (EBL) (Fig. 1). Details are provided in the Supplementary Notes and Figures. For any reasonable range of EBL fluxes at ultraviolet (UV) and optical/near-infrared wavelengths (O–NIR), $\tau(E)$ —and thus absorption—is larger at 1 TeV than at 0.2 TeV. This difference makes the observed spectrum steeper (that is, $F_{\text{obs}} > F_{\text{int}}$ for a power-law model $dN/dE \propto E^{-\Gamma}$). The spectral change $\Delta\Gamma = \Gamma_{\text{obs}} - \Gamma_{\text{int}}$ scales linearly with the EBL normalization, and becomes more pronounced at larger redshifts. Thus, more distant objects provide a more sensitive diagnostic tool.

In general, if the intrinsic spectrum were sufficiently well known, $\tau(E)$ —and thus the SED of the EBL—could be effectively measured by comparing intrinsic with observed spectra. Blazars, however, are characterized by a wide range of possible spectra, and the present understanding of their radiation processes is not yet complete enough to reliably predict their intrinsic γ -ray spectra. But for these two sources, with O–NIR fluxes at the level of the 'direct' estimates, the intrinsic spectra needed to reproduce the HESS data become extremely hard (that is, they have small values of Γ), at odds with the currently known blazar physics and phenomenology. This

¹Max-Planck-Institut für Kernphysik, PO Box 103980, D 69029 Heidelberg, Germany. ²Yerevan Physics Institute, 2 Alikhanian Brothers Street, 375036 Yerevan, Armenia.

³Centre d'Etude Spatiale des Rayonnements, CNRS/UPS, 9 avenue du Colonel Roche, BP 4346, F-31029 Toulouse cedex 4, France. ⁴Universität Hamburg, Institut für Experimentalphysik, Luruper Chaussee 149, D 22761 Hamburg, Germany. ⁵Institut für Physik, Humboldt-Universität zu Berlin, Newtonstrasse 15, D 12489 Berlin, Germany.

⁶LUTH, UMR 8102 du CNRS, Observatoire de Paris, Section de Meudon, F-92195 Meudon cedex, France. ⁷University of Durham, Department of Physics, South Road, Durham DH1 3LE, UK. ⁸Laboratoire Leprince-Ringuet, IN2P3/CNRS, Ecole Polytechnique, F-91128 Palaiseau, France. ⁹European Associated Laboratory for Gamma-Ray Astronomy.

¹⁰APC, 11 Place Marcelin Berthelot, F-75231 Paris cedex 05, France. ¹¹Dublin Institute for Advanced Studies, 5 Merrion Square, Dublin 2, Ireland. ¹²Landessternwarte, Königstuhl, D 69117 Heidelberg, Germany. ¹³Laboratoire de Physique Théorique et Astroparticules, IN2P3/CNRS, Université Montpellier II, CC 70, Place Eugène Bataillon, F-34095 Montpellier Cedex 5, France. ¹⁴Laboratoire d'Astrophysique de Grenoble, INSU/CNRS, Université Joseph Fourier, BP 53, F-38041 Grenoble cedex 9, France. ¹⁵DAPNIA/DSM/CEA, CE Saclay, F-91191 Gif-sur-Yvette cedex, France. ¹⁶Unit for Space Physics, North-West University, Potchefstroom 2520, South Africa. ¹⁷Laboratoire de Physique Nucléaire et de Hautes Energies, IN2P3/CNRS, Universités Paris VI & VII, 4 Place Jussieu, F-75252 Paris cedex 5, France. ¹⁸Institut für Theoretische Physik, Lehrstuhl IV: Weltraum und Astrophysik, Ruhr-Universität Bochum, D 44780 Bochum, Germany. ¹⁹Institute of Particle and Nuclear Physics, Charles University, V Holesovickach 2, 180 00 Prague 8, Czech Republic. ²⁰University of Namibia, Private Bag 13301, Windhoek, Namibia.

Table 1 | Main parameters of the HESS observations.

Source	z	Exposure (h)	Significance (σ)	Energy range (TeV)	Γ_{obs}	N_0 ($\text{cm}^{-2} \text{s}^{-1} \text{TeV}^{-1}$)	$\chi^2_{\text{red}}/\text{d.o.f.}$
1ES 101–232	0.186	43	~ 12	0.16–3.3	2.88 ± 0.17	$(4.44 \pm 0.74) \times 10^{-13}$	0.62/11
H 2356–309	0.165	40	~ 10	0.16–1.0	3.06 ± 0.21	$(3.08 \pm 0.75) \times 10^{-13}$	0.66/6

These observations were performed in June–December 2004 for H 2356–309, and March–June 2004 and 2005 for 1ES 101–232. The table gives the total exposure after selection for good-quality data, significance of the detected γ -ray signal, energy range used for the spectral fits and the result of a single power-law fit ($dN/dE = N_0 E^{-\Gamma_{\text{obs}}}$, where E is measured in TeV). The spectra have been calculated applying the technique described in ref. 10. Errors are 1σ statistical. The systematic uncertainty on the flux and photon index are estimated to be $\sim 15\%$ and ~ 0.1 , respectively. Details of these observations will be published elsewhere; here we have focused on the cosmological implications of the measured spectra. Compared to the previous observations of TeV blazars, these HESS spectra provide significantly stronger constraints on the EBL because of the combination of a hard spectrum and relatively high redshift (see Supplementary Information). d.o.f., degrees of freedom.

can be avoided by reducing the energy dependence of the optical depth, assuming either very low O–NIR fluxes (reducing τ at 1 TeV) or very high UV–O fluxes (increasing τ at 0.2 TeV). The latter case, however, would require unreasonably high UV fluxes (as discussed later). We can then derive an upper limit on the EBL by requiring the intrinsic spectrum to be compatible with the present knowledge of blazars.

To determine such a limit, a plausible shape for the SED of the EBL is assumed. As a reference shape in the 0.1–10 μm range we adopted the phenomenological curve used in refs 6 and 11, which is designed to be in general agreement with the EBL spectrum expected from galaxy emission^{12,13}. This curve, labelled ‘P1.0’ in Figs 1 and 2, was originally normalized to match the ‘direct’ estimates at 2.2 and 3.5 μm (refs 1, 14 and 15). Here we leave its normalization as a free parameter, scaling P1.0 by different factors P (labelled accordingly: the curve scaled by 0.45 $\times$ is ‘P0.45’) down to the lower limit obtained by the resolved galaxy counts⁸ ($\sim P0.4$). To reproduce the excess around 1.5 μm claimed from the Infrared Telescope in Space (IRTS) data¹⁶ (see also ref. 17), an additional *ad hoc* component was considered, labelled ‘ E_{NIR} ’. This feature is not expected from standard galaxy evolution models, and could be the spectral signature of radiation produced in the early Universe, for example, by the first stars formed (metal-free massive stars, called ‘population III’; see refs 9 and 18).

The intrinsic source spectra (Fig. 2) have been reconstructed directly from the observed ones using the assumed EBL, without a priori assumptions on the blazar spectrum. EBL evolution effects (due to galaxy evolution) were not included: these effects, negligible at low redshifts, become important as the redshift increases, but for the range considered here ($z = 0.165\text{--}0.186$) their impact is still limited to a factor of $\lesssim 10\%$ ($\Delta P < 0.1$; see Supplementary Information). The reconstructed spectra are generally compatible with a power law ($dN/dE \propto E^{-\Gamma}$), but the EBL densities P1.0 + E_{NIR} and P1.0 both yield extremely hard photon indices ($\Gamma_{\text{int}} < 0$, see Fig. 2 and Supplementary Fig. 2), implying a pile-up or line-like feature in the blazars’ SED at around 1–3 TeV. We obtain the same result by considering the NIR excess added to the galaxy count limits (for example, P0.4 + E_{NIR} , see Supplementary Fig. 3). This is because a lower EBL flux only in the UV–O band decreases the absolute values of τ but increases the contrast between 0.2 and 1 TeV.

Such hard spectra have never been seen in the closest, less absorbed TeV blazars such as Mkn 421 and Mkn 501 (refs 19–21) ($z \approx 0.03$, $\Gamma_{\text{int}} \approx 1.5\text{--}2.8$ using the same EBL SED values), and are difficult to explain within the present standard leptonic or hadronic scenarios⁶ for blazar emission. In shock acceleration models, the hardest index obtained for the accelerated particles is $s = 1.5$ (see ref. 22). For protons interacting with ambient plasma, the resulting γ -ray spectrum has the same slope, $\Gamma_{\text{int}} = 1.5$. For electrons, the spectrum of the γ -rays emitted through inverse Compton scattering is expected to be steeper than $\Gamma_{\text{int}} = 1.5$ under most circumstances. Only if radiative cooling is not effective and the blazar Compton emission is wholly within the Thomson limit—unlikely at such high energies—do we find $\Gamma_{\text{int}} = (s + 1)/2 = 1.25$. We thus assume in the following discussion that the true average intrinsic spectrum was not harder than $\Gamma_{\text{int}} = 1.5$, although later we also address the possibility of harder spectra.

To be compatible with $\Gamma_{\text{int}} \geq 1.5$, the EBL flux has to be scaled down to P0.45 to explain both objects’ data, with 1ES 1101–232 providing the most stringent constraints thanks to the better statistics at high energies and the larger redshift. With a fixed EBL shape, there is a direct link between the normalization P and Γ_{int} . The one-sigma statistical and systematic uncertainties on the HESS spectral measurement can then be translated to an equivalent uncertainty on P , $\Delta P \approx \pm 0.15$ (see Supplementary Information).

This limit (P0.45) is robust with respect to a different EBL spectral shape, as long as it maintains an overall maximum around 1–2 μm . Below 1 μm , lower fluxes than our template tend to harden the

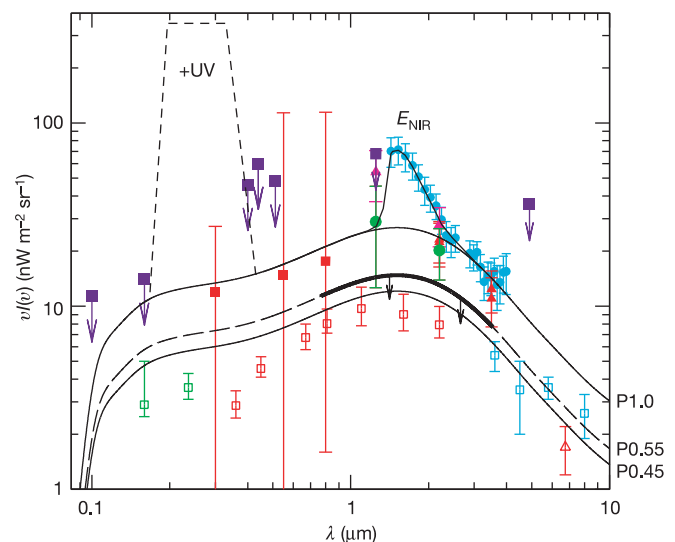


Figure 1 | SED of the EBL in the wavelength band most affecting these HESS data (0.1–10 μm). The EBL data are from a review compilation¹ (errors 1σ), unless otherwise stated. Open symbols correspond to the integrated light from galaxy counts, and thus must be considered lower limits for the EBL: in the UV–O range, from Hubble data (green, red⁸); in the NIR, from Spitzer (blue²⁸) and ISO data. We note that these data are also lower limits for the total emission from galaxies, because of various observational and selection effects in the detection and counting of faint galaxies. The possible missed light in the the UV–O band has been estimated²⁹ to be $\lesssim (2\text{--}3) \text{ nW m}^{-2} \text{ sr}^{-1}$. The upper limits (purple) are 2σ estimates. Direct measurements are shown as filled symbols: IRTS data from the NIR spectrometer¹⁶ (blue), and data from COBE/DIRBE (green¹⁵, magenta¹⁷ and red triangles). Red squares correspond to tentative detections in the optical¹⁶ with corrections according to ref. 30. The curves show the EBL shapes used to reconstruct the intrinsic spectra. P1.0 gives 26, 23 and 14 $\text{nW m}^{-2} \text{ sr}^{-1}$ at 1.25, 2.2 and 3.5 μm , respectively. The thick line shows the range most effectively constrained by the HESS data. In the long-dashed regions, higher fluxes than P0.55 would not be in conflict, as long as the flux in the 1–3 μm range is within or around the limit. The short dashed line shows the additional UV component needed by P1.0 to soften the intrinsic spectrum down to $\Gamma = 1.5$ (see Supplementary Fig. 5; E_{NIR} would require even higher fluxes). This example is the most energetically economic solution, limited to the narrow range $\sim 0.2\text{--}0.4 \mu\text{m}$ to have the maximum effect on the γ -ray spectrum with the minimum UV flux and the minimum impact on the overall attenuation.

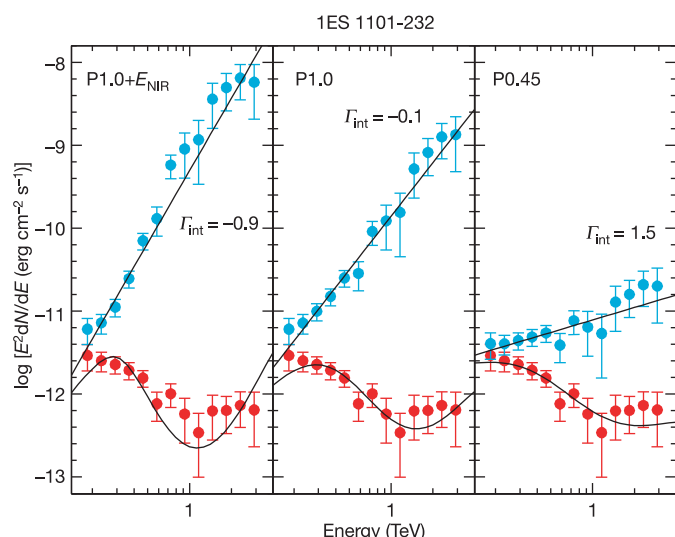


Figure 2 | The HESS spectra of 1ES 1101–232, corrected for absorption with three different EBL SED values, as labelled in Fig. 1. Red, observed data; blue, absorption-corrected data. The data points are at the average photon energy in each bin, also used to calculate the optical depth for reconstruction. For the calculation, a flat Λ -dominated cosmology was adopted, with $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$, $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$. Error bars are 1σ s.d., statistical errors only. Between 1.3 and 3.3 TeV, the overall detection is 4σ . The lines show the best-fit power laws to the reconstructed spectrum ($dN/dE = N_0 E^{-\Gamma_{\text{int}}}$), where E is measured in TeV, and the corresponding shapes after absorption. The $\chi^2_{\text{red}}/\text{d.o.f.}$ (calculated by integrating the absorbed power-law model over each observed data bin) are from left to right: 1.20/11, 0.54/11, 0.47/11. We note that possible spectral variability does not weaken our conclusions because it would imply states with even harder spectra than the average one (by definition). We note also that the X-ray spectrum (which in blazars usually samples the synchrotron emission of TeV electrons) measured during simultaneous observations in June 2004 and March 2005 does not show such hard slopes, but is similar to the historical values (F.A. *et al.*, manuscript in preparation). For H 2356–309, the same EBL SEDs yield $\Gamma_{\text{int}} = -0.6, 0.7$ and 2.0 , respectively (Supplementary Fig. 2). The NIR excess onto the galaxy counts limits ($P0.4 + E_{\text{NIR}}$) yields $\Gamma_{\text{int}} \approx -0.7$ and -0.4 for the two objects, respectively (see Supplementary Fig. 3).

intrinsic spectrum more, while even a flat slope from $1.4 \mu\text{m}$ down to $0.1 \mu\text{m}$ would soften it by only $\Delta\Gamma \approx 0.1$. Above $2 \mu\text{m}$, the slope cannot be much flatter than our template—a flatter slope could explain in part the ‘direct’ measurements at $3.5 \mu\text{m}$ (Supplementary Fig. 4)—because this would again imply a new, very hard component ($\Gamma_{\text{int}} < 0$) in the intrinsic spectrum, rising at a few TeV (Supplementary Fig. 5). In this respect, this HESS data set gives the same indication as the HEGRA data¹¹ on H 1426+428 ($z = 0.129$), which show a flattening feature above 1 TeV naturally provided by a starlight EBL between 3 and $10 \mu\text{m}$ ($\text{SED} \propto \lambda^{-1}$).

Therefore, the conservative and self-consistent assumptions of both not-unusual blazar spectra ($\Gamma_{\text{int}} \geq 1.5$) and a galaxy-like EBL spectrum allow the EBL flux around $1\text{--}2 \mu\text{m}$ to be constrained at the level of $\leq (14 \pm 4) \text{ nW m}^{-2} \text{ sr}^{-1}$ (that is, $\leq 0.55 \pm 0.15 \times P1.0$). This corresponds to $P(0.45 \pm 0.1)$ to allow for galaxy evolution effects. Coupled with the lower limits derived from galaxy counts given by the Hubble Space Telescope⁸ ($\sim 9.0 - 9.7^{+3.0}_{-1.9} \text{ nW m}^{-2} \text{ sr}^{-1}$), the HESS spectra lead us to conclude that more than two-thirds of the EBL in the O–NIR band is resolved into single sources. This result is independent of any ‘direct’ measurement of the EBL. Remarkably, it is in severe conflict with the claims of high EBL flux at NIR wavelengths^{16,17} and, to a lesser extent, with the reported detections at 2.2 and $3.5 \mu\text{m}$ (refs 1, 15). The HESS upper limits agree instead with the most recent theoretical calculations²³ of the EBL, as well as with recent theoretical arguments^{24,25} against high EBL fluxes due to population III stars.

This result is also insensitive to small changes in the assumed intrinsic slope. A different value, if proved more likely by future results, will shift the limit accordingly, but only strong spectral differences would qualitatively change our conclusion: even a value of $\Gamma_{\text{int}} = 1.0$ would loosen the $0.55\times$ limit only to $\leq 0.7 \times P1.0$.

Alternative scenarios which could reconcile the measured spectra with high O–NIR fluxes formally exist, and would represent a major discovery in their own right, but we consider them very unlikely, given their exotic implications. Higher UV fluxes would make the intrinsic spectra softer, but the huge values required ($> 300 \text{ nW m}^{-2} \text{ sr}^{-1}$; see Fig. 1, for example) are not supported by other measurements^{1,26}, and could not be accommodated within any reasonable cosmological model.

A more viable alternative is that such hard spectra are a real, new feature of the TeV blazar emission. Possible mechanisms have already been envisaged⁶. For example, the inverse Compton scattering of mono-energetic electrons (E_0 , such as cold plasma with very large bulk motion Lorentz factor), interacting in a deep Klein–Nishina regime with a narrow-band photon field (such as a Planck-type distribution), may lead to very flat γ -ray spectra with a sharp pile-up at $\epsilon_\gamma \approx E_0$, reproducing spectra like the ones in Fig. 2. However, such features should become directly visible in the observed spectra of the closer, less absorbed objects of the same type, like the well-studied Mkn 421 and Mkn 501 (if $\Gamma_{\text{int}} \approx 0$, they should show $\Gamma_{\text{obs}} \lesssim 0.5$). This is in contrast with observations^{19–21}, unless we assume a dependence of the source parameters on the redshift such that the corresponding features always disappear owing to EBL absorption. It is difficult to justify such fine-tuning on a relatively small redshift range, although more objects and observations are needed to settle this issue definitively, given the still-limited sample.

Other possibilities include the non-cosmological origin of blazars’ redshifts and the violation of Lorentz invariance (see ref. 27). However, both scenarios imply dramatic revisions of modern physics and astrophysics, which we do not consider to be justified by these data alone.

A low EBL level, in agreement with the expectations from standard galaxy evolution models, is the simplest and most likely explanation of the HESS data.

Received 25 July 2005; accepted 22 February 2006.

- Hauser, M. G. & Dwek, E. The Cosmic Infrared Background: measurements and implications. *Annu. Rev. Astron. Astrophys.* **39**, 249–307 (2001).
- Nikishov, A. I. Absorption of high-energy photons in the universe. *Sov. Phys. JETP* **14**, 393–394 (1962).
- Gould, R. J. & Schreder, G. P. Opacity of the Universe to high-energy photons. *Phys. Rev. Lett.* **16**, 252–254 (1966).
- Jelley, J. V. High-energy gamma-ray absorption in space by a 3.5°K microwave field. *Phys. Rev. Lett.* **16**, 479–481 (1966).
- Stecker, F. W., De Jager, O. C. & Salamon, M. H. TeV gamma rays from 3C 279—A possible probe of origin and intergalactic infrared radiation fields. *Astrophys. J. Lett.* **390**, 49–52 (1992).
- Aharonian, F. A. TeV blazars and the cosmic infrared background radiation. *Invited, Rapporteur, and Highlight Papers of the Proc. 27th ICRC (Hamburg)* (ed. Schlickeiser, R.) 250–262 (Copernicus Gesellschaft, Katlenburg-Lindau, 2001); preprint at (<http://arXiv.org/astro-ph/0112314>) (2001).
- Costamante, L. & Ghisellini, G. TeV candidate BL Lac objects. *Astron. Astrophys.* **384**, 56–71 (2002).
- Madau, P. & Pozzetti, L. Deep galaxy counts, extragalactic background light and the stellar baryon budget. *Mon. Not. R. Astron. Soc.* **312L**, 9–15 (2000).
- Santos, M. R., Bromm, V. & Kamionkowski, M. The contribution of the first stars to the cosmic infrared background. *Mon. Not. R. Astron. Soc.* **336**, 1082–1092 (2002).
- Aharonian, F. A. *et al.* H.E.S.S. observations of PKS 2155–304. *Astron. Astrophys.* **430**, 865–875 (2005).
- Aharonian, F. A. *et al.* Observations of H1426+428 with HEGRA. Observations in 2002 and reanalysis of 1999 and 2000 data. *Astron. Astrophys.* **403**, 523–528 (2003).
- Primack, J. R., Somerville, R. S., Bullock, J. S. & Devriendt, J. E. G. Probing Galaxy formation with high-energy gamma rays. *AIP Conf. Proc.* **558**, 463–478 (2001).
- Dwek, E. & Krennrich, F. Simultaneous constraints on the spectrum of the extragalactic background light and the intrinsic TeV spectra of Mkn 421, Mkn 501, and H1426+428. *Astrophys. J.* **618**, 657–674 (2005).

14. Dwek, E. & Arendt, R. G. A tentative detection of the cosmic infrared background at $3.5\text{ }\mu\text{m}$ from COBE/DIRBE observations. *Astrophys. J. Lett.* **508**, 9–12 (1998).
15. Wright, E. L. DIRBE minus 2MASS: confirming the cosmic infrared background at 2.2 microns. *Astrophys. J.* **553**, 538–544 (2001).
16. Matsumoto, T. *et al.* Infrared telescope in space observations of the near-infrared extragalactic background light. *Astrophys. J.* **626**, 31–43 (2005).
17. Cambresy, L., Reach, W. T., Beichman, C. A. & Jarrett, T. H. The cosmic infrared background at 1.25 and 2.2 microns using DIRBE and 2MASS: A contribution not due to galaxies? *Astrophys. J.* **555**, 563–571 (2001).
18. Kashlinsky, A. Cosmic infrared background and early galaxy evolution. *Phys. Rep.* **409**, 361–438 (2005).
19. Aharonian, F. *et al.* The time averaged TeV energy spectrum of MKN 501 of the extraordinary 1997 outburst as measured with the stereoscopic Cherenkov telescope system of HEGRA. *Astron. Astrophys.* **349**, 11–28 (1999).
20. Krennrich, F. *et al.* Discovery of spectral variability of Markarian 421 at TeV energies. *Astrophys. J. Lett.* **575**, 9–13 (2002).
21. Djannati-Atai, A. *et al.* Very high energy gamma-ray spectral properties of MKN 501 from CAT Cherenkov telescope observations in 1997. *Astron. Astrophys.* **350**, 17–24 (1999).
22. Malkov, M. A. & Drury, L. O'C. Nonlinear theory of diffusive acceleration of particles by shock waves. *Rep. Prog. Phys.* **64**, 429–481 (2001).
23. Primack, J. R., Bullock, J. S. & Somerville, R. S. Observational gamma-ray cosmology. *AIP Conf. Proc.* **745**, 23–33 (2005).
24. Madau, P. & Silk, J. Population III and the near-infrared background excess. *Mon. Not. R. Astron. Soc.* **359L**, 37–41 (2005).
25. Dwek, E., Arendt, R. G. & Krennrich, F. The near infrared background: interplanetary dust or primordial stars? *Astrophys. J.* **635**, 784–794 (2005).
26. Bernstein, R. A., Freedman, W. L. & Madore, B. F. The first detections of the extragalactic background light at 3000, 5500, and 8000 Å. I. Results. *Astrophys. J.* **571**, 56–84 (2002).
27. Protheroe, R. J. & Meyer, H. An infrared background-TeV gamma-ray crisis? *Phys. Lett. B* **493**, 1–6 (2000).
28. Fazio, G. G. *et al.* Number counts at $3\text{ }\mu\text{m} < \lambda < 10\text{ }\mu\text{m}$ from the Spitzer Space Telescope. *Astrophys. J. Suppl.* **154**, 39–43 (2004).
29. Totani, T. *et al.* Diffuse extragalactic background light versus deep galaxy counts in the Subaru deep field: Missing light in the Universe? *Astrophys. J.* **550L**, 137–141 (2001).
30. Mattila, K. Has the optical extragalactic background light been detected? *Astrophys. J.* **591**, 119–124 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The support of the Namibian authorities and of the University of Namibia in facilitating the construction and operation of HESS is gratefully acknowledged, as is the support by the German Ministry for Education and Research (BMBF), the Max Planck Society, the French Ministry for Research, the CNRS-IN2P3 and the Astroparticle Interdisciplinary Programme of the CNRS, the UK Particle Physics and Astronomy Research Council (PPARC), the IPNP of the Charles University, the South African Department of Science and Technology and National Research Foundation, and by the University of Namibia. We appreciate the excellent work of the technical support staff in Berlin, Durham, Hamburg, Heidelberg, Palaiseau, Paris, Saclay, and in Namibia in the construction and operation of the equipment. The European Associated Laboratory for Gamma-Ray Astronomy is jointly supported by CNRS and MPG.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to L.C. (luigi.costamante@mpi-hd.mpg.de).

LETTERS

Experimental determination of entanglement with a single measurement

S. P. Walborn¹, P. H. Souto Ribeiro¹, L. Davidovich¹, F. Mintert^{1,2} & A. Buchleitner³

Nearly all protocols requiring shared quantum information¹—such as quantum teleportation² or key distribution³—rely on entanglement between distant parties. However, entanglement is difficult to characterize experimentally. All existing techniques for doing so, including entanglement witnesses^{4,11,12} or Bell inequalities⁵, disclose the entanglement of some quantum states but fail for other states; therefore, they cannot provide satisfactory results in general. Such methods are fundamentally different from entanglement measures that, by definition, quantify the amount of entanglement in any state. However, these measures suffer from the severe disadvantage that they typically are not directly accessible in laboratory experiments. Here we report a linear optics experiment in which we directly observe a pure-state entanglement measure, namely concurrence⁶. Our measurement set-up includes two copies of a quantum state: these ‘twin’ states are prepared in the polarization and momentum degrees of freedom of two photons, and concurrence is measured with a single, local measurement on just one of the photons.

The functional relation between an entanglement measure and the state to be characterized is typically complicated and nonlinear. Even worse, many measures rely on operations that cannot be implemented in laboratory experiments for fundamental reasons: they do not preserve the positivity of general quantum states. That is, these operations would spoil the statistical interpretation of quantum mechanics and thus cannot be realized in any physical system. In particular, this holds for the most commonly used measures, namely concurrence⁷ and negativity⁸. Thus, it is necessary to measure a complete set of observables, reconstruct the system state, and eventually evaluate them. This—although possible and successfully implemented^{9,10} for relatively small systems—becomes virtually impossible with larger systems because the number of observables to be measured grows exponentially with the number of entangled subsystems. On the other hand, one might expect that a property like entanglement, which is believed to constitute one of the most remarkable differences between classical and quantum systems, should have signatures that can be observed directly. And indeed, there is a proposal¹³ to directly observe concurrence by replacing the underlying unphysical operation by some physical approximation. This, however, requires a rather involved experimental set-up, and has not been realized yet.

Here, we report the direct experimental observation of an entanglement measure, namely concurrence⁷

$$C = |\langle \Psi^* | \sigma_y \otimes \sigma_y | \Psi \rangle|, \quad \sigma_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \quad (1)$$

originally defined in terms of the second Pauli matrix σ_y and the complex conjugate $\langle \Psi^* |$ of the original state $\langle \Psi |$. The problems that arise owing to the unphysical operation of complex conjugation in equation (1) are overcome here with a generalization of concurrence¹⁴

to systems of arbitrary finite dimensions, which, if restricted to qubits, is equivalent to the original concurrence. In this case the nonlinear dependence on the system state—that constitutes a fundamental property of any entanglement measure—is taken into account by considering a twofold copy of the state in question. Indeed, it has been shown that any m th degree polynomial function of a density matrix ρ can be measured on an m -fold copy of ρ (ref. 15). More precisely, the concurrence C of an arbitrary state $|\Psi\rangle$ can be defined as $C = 2\sqrt{P_A}$, where $P_A = \langle \Psi | \otimes \langle \Psi | A | \Psi \rangle \otimes | \Psi \rangle$ is the probability of observing the two copies of the first subsystem in an antisymmetric state, that is, a state that acquires a phase shift of π upon exchange of the constituents, and A is the corresponding measurement operator. In particular, no measurement needs to be performed on either copy of the second subsystem—though, in general, detections on the second subsystem can be used to trigger detectors for measurements on the first subsystem.

In our specific set-up, shown in Fig. 1, we created two copies of a bipartite quantum state using a photon pair obtained by spontaneous parametric down-conversion, where the polarization and momentum degrees of freedom each store one copy of the state $|\Psi\rangle$,

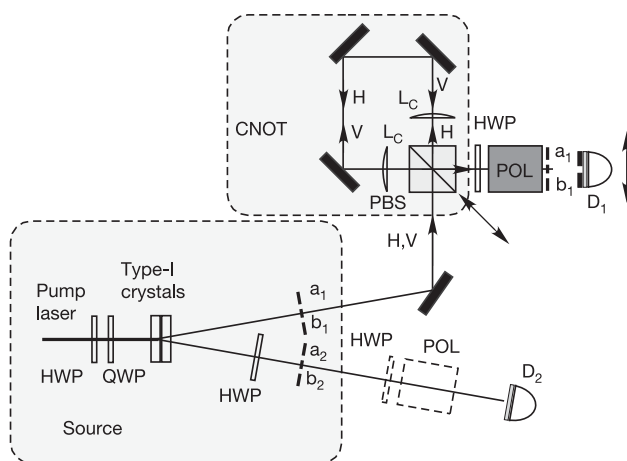


Figure 1 | Experimental set-up for the measurement of entanglement using two copies of the quantum state. Photon pairs that bear entanglement in two different degrees of freedom were created by pumping two type-I LiIO₃ crystals with a 200 mW HeCd continuous-wave laser (442 nm). Double-square apertures (1 mm × 1 mm squares, 2 mm centre to centre separation) placed 1 m from the crystal face are used to select distinct spatial modes *a* and *b*. Detectors D₁ and D₂ use 1.4 mm circular and 1 × 5 mm rectangular detection apertures, respectively. Both were equipped with interference filters (full-width at half-maximum, 10 nm). HWP, half-wave plate; QWP, quarter-wave plate; PBS, polarizing beam splitter; L_c, cylindrical lens; POL, polarization filter. CNOT, controlled-not. H and V indicate horizontal and vertical polarization.

¹Instituto de Física, Universidade Federal do Rio de Janeiro, Caixa Postal 68528, Rio de Janeiro, RJ 21941-972, Brazil. ²Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA. ³Max-Planck-Institut für Physik komplexer Systeme, Nöthnitzer Strasse 38, D-01187 Dresden, Germany.

respectively^{16,17}. That is, both copies required for our measurement are carried by the same photon, which significantly facilitates the set-up, as a measurement on only a single photon will be necessary. In fact, any required measurement can be realized efficiently with linear optics only^{18,19}.

Entangled polarization states were prepared by pumping two perpendicular nonlinear crystals with a continuous-wave laser beam²⁰. The additional help of half- and quarter-wave plates placed in the path of both the pump beam and the entangled photons eventually allows us to prepare arbitrary pure polarization states. Moreover, owing to momentum conservation, entanglement between the momentum degrees of freedom can also be achieved using apertures to select well-defined momentum modes^{16,21}; and the whole range of momentum states can be generated with neutral filters, phase plates and beam splitters that combine different momentum modes¹⁷.

With these two degrees of freedom, the entire system of two photons has polarization states spanned by $|H\rangle_i$, $|V\rangle_i$ ($i = 1, 2$ labels the two photons) and momentum states spanned by $|a\rangle_i$ and $|b\rangle_i$. Upon identification of the momentum state $|a\rangle_i$ as the equivalent of the polarization state $|H\rangle_i$ and analogously for $|b\rangle_i$ and $|V\rangle_i$, one can prepare two copies of an arbitrary input state $|\Psi\rangle$, one stored in each degree of freedom.

As outlined above, the concurrence of $|\Psi\rangle$ is determined by the probability of observing the first photon in the antisymmetric state

$$|\psi^-\rangle = \frac{1}{\sqrt{2}}(|H\rangle|b\rangle - |V\rangle|a\rangle) \quad (2)$$

where we have dropped the index '1', as from now on all considerations will concern only the first photon. As count rates rather than probabilities are accessible in laboratory experiments, we also need to count events corresponding to the detection of the remaining Bell states:

$$|\psi^+\rangle = \frac{1}{\sqrt{2}}(|H\rangle|b\rangle + |V\rangle|a\rangle) \quad (3a)$$

$$|\phi^\pm\rangle = \frac{1}{\sqrt{2}}(|H\rangle|a\rangle \pm |V\rangle|b\rangle) \quad (3b)$$

The probability P_A , which determines concurrence, is then given by the count rate for the observation of $|\psi^-\rangle$ normalized by the sum of the count rates for all four Bell states.

The central building block for this Bell-state measurement in our specific experimental set-up was a polarization-sensitive Sagnac

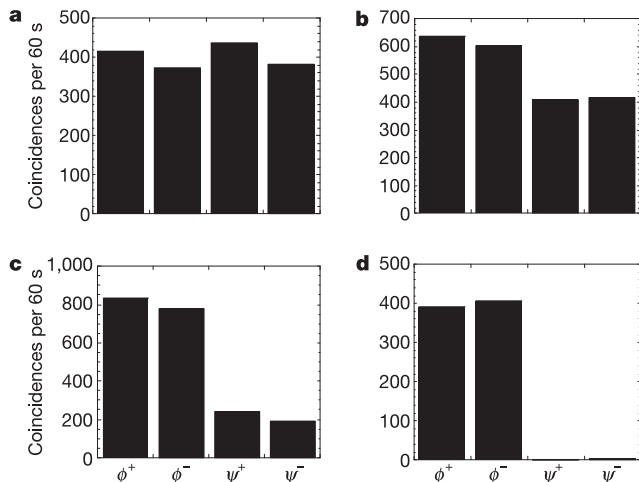


Figure 2 | Count rates. Experimentally obtained count rates of the Bell-state measurement (see equations (2) and (3)) on the twofold copy of input states $\alpha|01\rangle + \beta|10\rangle$ with **a**, $|\alpha| = 0.71 \pm 0.02$, **b**, $|\alpha| = 0.53 \pm 0.01$, **c**, $|\alpha| = 0.35 \pm 0.01$, and **d**, $|\alpha| = 0.99 \pm 0.03$.

interferometer containing two cylindrical lenses, as depicted in Fig. 1. The interferometer is used to perform a polarization-dependent rotation of the momentum modes, which is equivalent to a controlled-not (CNOT) operation. Input photons are first incident on a polarizing beam splitter, which transmits H-polarized photons and reflects V-polarized photons, so that H- and V-polarized photons propagate in opposite directions within the interferometer, and leave through the same exit port. An optical lens system, consisting of two 150 mm focal length cylindrical lenses, was rotated by 45° in the transverse plane with respect to the propagation direction of the H-polarized photons. The individual lenses were aligned at $+45^\circ$ and -45° . As each photon suffered three mirror reflections while propagating between the lenses, it was necessary to invert the orientation of one of them. The lenses were placed in a confocal arrangement, so that each lens was a distance of 150 mm from the double square aperture, and also 150 mm from the central mirror of the interferometer, forming a confocal imaging system with a magnification factor of one. The image formed by a cylindrical lens is inverted with respect to one transverse direction only, so that a lens aligned at $\pm 45^\circ$ forms an image that is rotated $\pm 90^\circ$ with respect to the object. Because H- and V-polarized photons counter-propagate within the interferometer, they encounter the lens oriented at different angles ($\pm 45^\circ$), so that the resulting image corresponding to H-polarized photons is rotated by 90° , while the image corresponding to the V-polarized photons is rotated by -90° , resulting in a relative difference of 180° . In the image plane, the Sagnac interferometer with the nested cylindrical lenses performs the so-called CNOT operation, where the momentum state evolves conditioned on the polarization—if the photon is vertically polarized, $|a\rangle$ evolves to $|b\rangle$ and vice versa; otherwise the momentum states remain unchanged. In reality, both the momentum states encounter an additional rotation by 90° , which can easily be accounted for by simply rotating our coordinate system. The operation of a similar CNOT gate was characterized in ref. 22. The crucial benefit of the CNOT operation is that it transforms the Bell states such that the momentum and polarization states factorize:

$$\text{CNOT}(|\psi^\pm\rangle) = 1/\sqrt{2}(|H\rangle \pm |V\rangle)|b\rangle = |\pm\rangle|b\rangle \quad (4a)$$

$$\text{CNOT}(|\phi^\pm\rangle) = 1/\sqrt{2}(|H\rangle \pm |V\rangle)|a\rangle = |\pm\rangle|a\rangle \quad (4b)$$

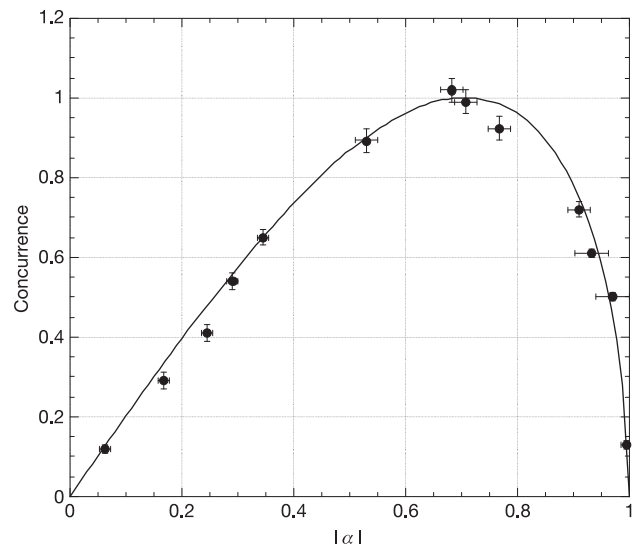


Figure 3 | Experimental values of concurrence. Data points, directly measured concurrence for states $\alpha|01\rangle + \beta|10\rangle$ as function of $|\alpha|$ with error bars due to poissonian statistics. The excellent agreement with the theoretical value $C = 2|\alpha|\sqrt{1-|\alpha|^2}$ (shown as the solid line) confirms the precision of the described measurement set-up.

Thus, observing a photon with $|-\rangle$ polarization and momentum $|b\rangle$ after the CNOT is equivalent to observing the state $|\psi^-\rangle$ before the CNOT operation, and analogously for the other Bell states. Thus, the final measurement simply consists of detecting $|\pm\rangle$ polarized photons in the modes a and b. This can easily be carried out with two detectors positioned in the paths of modes a and b, and additional half-wave plates and polarization analysers to discriminate the different polarizations. In particular, it should be emphasized that the four possible measurement results are the outcomes of a single measurement only.

In our experiment we measured the concurrence of states $\alpha|01\rangle + \beta|10\rangle$ with varying coefficients α and β . These states are particularly well suited, as the entire measurement protocol works perfectly, even for imperfect copies with different relative phases—which significantly relaxes the precision required during the preparation process. The coefficients α and β of the polarization and momentum degrees of freedom were varied by rotating the half-wave plate in the pump beam and shifting the apertures defining the momentum modes of photon 2, respectively. Figure 2 shows the experimental count rates for observations of the Bell states as defined in equations (2) and (3) for four different states with decreasing entanglement from Fig. 2a to Fig. 2d. Experimentally obtained concurrence C is depicted in Fig. 3 as a function of the varying coefficient α . The black dots show the experimentally obtained values, with error bars due to poissonian count statistics. The theoretical value of $C = 2|\alpha\beta| = 2|\alpha|\sqrt{1 - |\alpha|^2}$ is plotted as a solid line and agrees virtually perfectly with the experimental observations. In particular, the maximum value $C = 1$ is obtained for $|\alpha| = 1/\sqrt{2}$, which provides additional experimental evidence for the purity of the input states.

Our work shows that it is possible to directly assess entanglement properties with few—in this case only one—local measurements. Whereas state reconstruction and subsequent mathematical determination of entanglement is a viable and successfully demonstrated option for systems with few constituents, more efficient approaches are required for large objects. Our present experiment gives a proof of principle that indeed it is possible to circumvent the highly inefficient state reconstruction, and reliably characterize the entanglement properties of an unknown quantum state.

Received 6 December 2005; accepted 27 January 2006.

1. Nielsen, M. & Chuang, I. *Quantum Computation and Quantum Information* (Cambridge Univ. Press, Cambridge, UK, 2000).
2. Bennett, C. H. et al. Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Phys. Rev. Lett.* **70**, 1895–1899 (1993).
3. Ekert, A. K. Quantum cryptography based on Bells theorem. *Phys. Rev. Lett.* **67**, 661–663 (1991).

4. Horodecki, M., Horodecki, P. & Horodecki, R. Separability of mixed states: necessary and sufficient conditions. *Phys. Lett. A* **223**, 1–8 (1996).
5. Bell, J. S. On the Einstein Podolsky Rosen paradox. *Physics* **1**, 195–200 (1964).
6. Bennet, C. H., DiVincenzo, D. P., Smolin, J. A. & Wootters, W. K. Mixed-state entanglement and quantum error correction. *Phys. Rev. A* **54**, 3824–3851 (1996).
7. Wootters, W. K. Entanglement of formation of an arbitrary state of two qubits. *Phys. Rev. Lett.* **80**, 2245–2248 (1998).
8. Vidal, G. & Werner, R. F. Computable measure of entanglement. *Phys. Rev. A* **65**, 032314 (2002).
9. Sancho, J. M. G. & Huelga, S. F. Measuring the entanglement of bipartite pure states. *Phys. Rev. A* **61**, 042303 (2000).
10. White, A., James, D., Eberhard, P. & Kwiat, P. Nonmaximally entangled states: Production, characterization, and utilization. *Phys. Rev. Lett.* **83**, 3103–3107 (1990).
11. Häffner, H. et al. Scalable multiparticle entanglement of trapped ions. *Nature* **438**, 643–646 (2005).
12. Leibfried, D. et al. Creation of a six-atom 'Schrödinger cat' state. *Nature* **438**, 639–642 (2005).
13. Horodecki, P. Measuring quantum entanglement without prior state reconstruction. *Phys. Rev. Lett.* **90**, 167901 (2003).
14. Mintert, F., Kuš, M. & Buchleitner, A. Concurrence of mixed multi-partite quantum states. *Phys. Rev. Lett.* **95**, 260502 (2005).
15. Brun, T. A. Measuring polynomial functions of states. *Quant. Inform. Comput.* **4**, 401–408 (2004).
16. Kwiat, P. G. Hyper entangled states. *J. Mod. Opt.* **44**, 2173–2184 (1997).
17. Barbieri, M., Cinelli, C., Mataloni, P. & Martini, F. D. Polarization-momentum hyperentangled states: Realization and characterization. *Phys. Rev. A* **72**, 052110 (2005).
18. Boschi, D., Branca, S., DeMartini, F., Hardy, L. & Popescu, S. Experimental realization of teleporting an unknown pure quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Phys. Rev. Lett.* **80**, 1121–1125 (1998).
19. Englert, B.-G., Kurtsiefer, C. & Weinfurter, H. Universal unitary gate for single-photon two-qubit states. *Phys. Rev. A* **63**, 032303 (2003).
20. Kwiat, P. G., Waks, E., White, A. G., Appelbaum, I. & Eberhard, P. H. Ultrabright source of polarization-entangled photons. *Phys. Rev. A* **60**, R773–R776 (1999).
21. Rarity, J. & Tapster, P. Experimental violation of Bell's inequality based on phase and momentum. *Phys. Rev. Lett.* **64**, 2495–2498 (1990).
22. Fiorentino, M. & Wong, F. N. C. Deterministic controlled-not gate for single-photon two-qubit quantum logic. *Phys. Rev. Lett.* **93**, 070502 (2004).

Acknowledgements We acknowledge support from the Brazilian agencies CNPq, CAPES, PRONEX, FUIB and FAPERJ. This work was performed as part of the Brazilian Millennium Institute for Quantum Information, and was supported by a fellowship within the postdoctoral programme of the German Academic Exchange Service (DAAD) as well as by a Feodor Lynen fellowship of the Alexander von Humboldt foundation. We are indebted to C. H. Monkner for discussions.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.P.W. (swalborn@if.ufrj.br).

Polarons and confinement of electronic motion to two dimensions in a layered manganite

H. M. Rønnow¹, Ch. Renner², G. Aeppli², T. Kimura³ & Y. Tokura⁴

A remarkable feature of layered transition-metal oxides—most famously, the high-temperature superconductors—is that they can display hugely anisotropic electrical and optical properties (for example, seeming to be insulating perpendicular to the layers and metallic within them), even when prepared as bulk three-dimensional single crystals. This is the phenomenon of ‘confinement’, a concept at odds with the conventional theory of solids, and recognized¹ as due to magnetic and electron–lattice interactions within the layers that must be overcome at a substantial energy cost if electrons are to be transferred between layers. The associated energy gap, or ‘pseudogap’, is particularly obvious in experiments where charge is moved perpendicular to the planes, most notably scanning tunnelling microscopy² and polarized infrared spectroscopy³. Here, using the same experimental tools, we show that there is a second family of transition-metal oxides—the layered manganites $\text{La}_{2-2x}\text{Sr}_{1+2x}\text{Mn}_2\text{O}_7$ —with even more extreme confinement and pseudogap effects. The data demonstrate quantitatively that because the charge carriers are attached to polarons (lattice- and spin-textures within the planes), it is as difficult to remove them from the planes through vacuum-tunnelling into a conventional metallic tip, as it is for them to move between Mn-rich layers within the material itself.

The manganese perovskite oxides (manganites) have been the subject of a decade-long revival, spurred by the colossal magnetoresistance^{4,5} of thin films and sustained by the large variety of physical phenomena—including Jahn–Teller distortions, charge, spin and orbital ordering^{6–9}—manifested in carefully prepared single crystals. A fruitful new area has been the layered manganites, the most prominent of which is $\text{La}_{2-2x}\text{Sr}_{1+2x}\text{Mn}_2\text{O}_7$ (LSMO). They are built from perovskite bilayers as found in the pseudocubic parent compound $(\text{La,Sr})\text{MnO}_3$, separated by rock-salt-type lanthanide and alkaline-earth ion oxide sheets in a tetragonal crystal structure (Fig. 1a). These materials can be viewed as self-assembled stacks of two-dimensional electronic liquids weakly coupled to each other. They display the same anisotropic electric resistance (Fig. 1b) as the high-temperature superconductors (HTSs), and the same tunnelling magnetoresistance as top-down fabricated spintronic devices, where switching from the low- to the high-resistance state occurs via conversion from antiparallel to parallel configurations of the spins in adjacent bilayers¹⁰.

The lamellar structure of some HTS copper oxides has been very advantageous for experiments that are traditionally suitable for surface science, namely photoemission and scanning tunnelling microscopy (STM). Much less information of this type is available for the manganites, especially their layered polymorphs. Here we report atomic scale microscopy and temperature-dependent spectroscopy of unprecedentedly large and (for manganites) flat surfaces of LSMO, providing definitive evidence for confinement

but finding no evidence—in this particular oxide—for the popular concept of electronic phase separation¹¹ where different electronic states coexist in adjacent regions.

As a function of Sr doping x , LSMO exhibits the familiar features of many manganites¹². On cooling from the high-temperature paramagnetic state, LSMO develops a charge-ordered insulator around $x = 0.5$ and a variety of ferro- and antiferromagnetic phases between $x = 0.45$ and $x = 0.30$. We investigate the $x = 0.30$ compound,

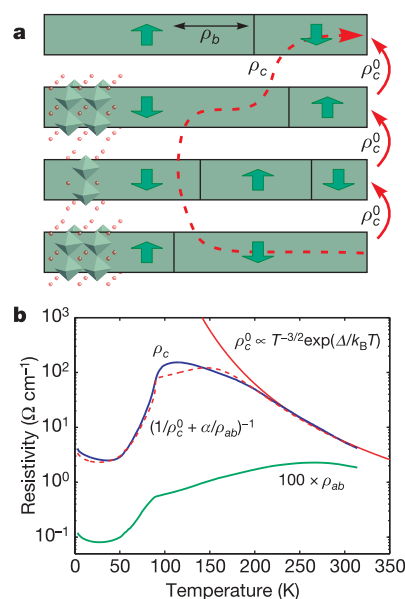


Figure 1 | Structure and bulk transport properties of $\text{La}_{1.4}\text{Sr}_{1.6}\text{Mn}_2\text{O}_7$.

a, Bilayer perovskite structure and different paths considered for the c -axis transport. Green arrows depict the ferromagnetic domains, and the dashed red line a possible reduced-resistance c -axis conduction path. **b**, The measured in-plane resistivity, ρ_{ab} (green), and c -axis resistivity, ρ_c (blue), drop dramatically below the ferromagnetic ordering temperature. The high-temperature part of the c -axis resistivity fits a purely activated behaviour (red), $\rho_c^0(T) \propto T^{-3/2} \exp(\Delta_p/2k_B T)$, with $\Delta_p = 188$ meV. The dashed red line depicts a model that includes in-plane contributions through conductivity enhancements at magnetic stacking faults. It places the intrinsic c -axis resistance $\rho_c^0(T)$ in parallel with a shunt resistance that is the product of $\rho_{ab}(T)$ and a prefactor $\alpha^{-1}(T)$, determined by the distances that the carriers travel within the planes, as well as the probability of moving carriers between parallel magnetized bilayers in the defective regions where adjacent bilayers have parallel magnetizations. Because the latter quantity grows with the magnetic order parameter M appearing below T_c , we use the Ansatz that $\alpha(T) = aM^2(T) + b$.

¹Laboratory for Neutron Scattering, ETH–Zürich and Paul Scherrer Institut, 5232 Villigen, Switzerland. ²London Centre for Nanotechnology & Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK. ³Bell Laboratories, Lucent Technologies, 600 Mountain Avenue, Murray Hill, New Jersey 07974, USA. ⁴Department of Applied Physics, University of Tokyo, Bunkyo-ku, Tokyo 113-8656, Japan, and Spin Superstructure Project (SSS), ERATO, Japan Science and Technology Agency (JST), Tsukuba 305-0046, Japan.

whose outstanding feature, shared with optimally and underdoped HTSs, is that in the paramagnetic phase the electrical resistance within the planes appears metallic—decreasing with decreasing temperature—while that corresponding to hopping between them is much larger and insulating (Fig. 1b). Below $T_c = 90$ K, the material orders as an antiferromagnetic stack of two-dimensional ferromagnets (the bilayers), and the electrical properties in zero field track each other, following the paradigm of dirty metals with quantum corrections¹³.

We performed STM using tungsten tips in ultrahigh vacuum at temperatures from 5 to 400 K. *In situ* cleavage at room temperature of the LSMO single crystals¹⁴ produced clean and extremely flat surfaces (peak-to-peak roughness below 0.3 Å over micrometre-sized terraces, Fig. 2a). As three-dimensional perovskites like LaMnO_3 do not cleave, LSMO must break between the perovskite bilayers, in agreement with the observed steps that are half a unit cell high (Fig. 2a). Yet, despite the smooth surfaces, atomic resolution remains elusive, except for tiny islands containing approximately 30 unit cells with the expected 3.86 Å Mn ion spacing lattice parameter (Fig. 2b), and slightly raised (above the background terrace) by an amount too small to be ascribed to adsorbed impurities. The low density of small and stable (reproducible over many hours, Fig. 2c–e) islands should not be regarded as phase separation, but as a local perturbation by some atomic-scale defect in the otherwise homogeneous system.

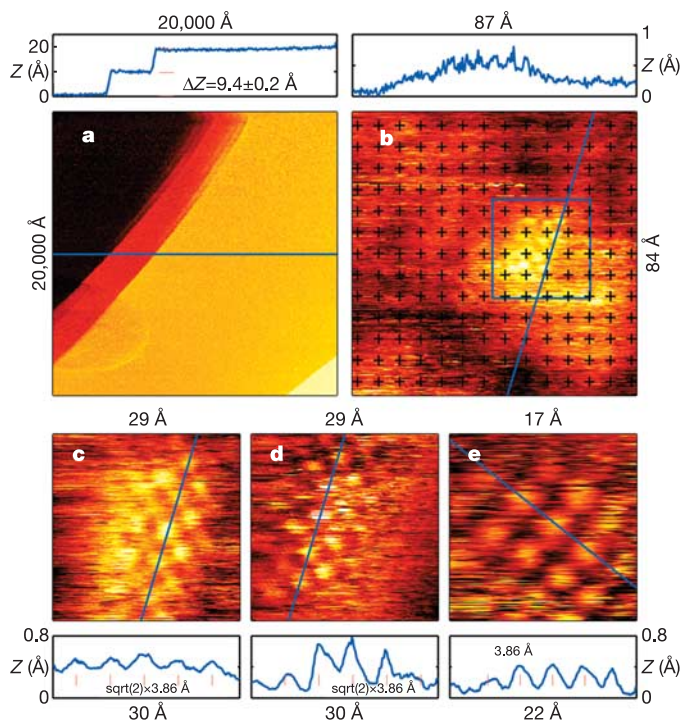


Figure 2 | Unprocessed STM micrographs of $\text{La}_{1.4}\text{Sr}_{1.6}\text{Mn}_2\text{O}_7$. The sample is held at a positive potential relative to the STM tip ($V = 0.8$ V, $I = 0.2$ nA), implying that we are primarily probing the unoccupied electronic density of states of the manganite. **a**, *In situ* cleaving exposed atomically flat micrometre-sized terraces (imaged at 212 K). On this scale, the nonlinear response of the piezoelectric STM scanner accounts for the curved appearance of the steps. Nominal scan amplitudes are given next to each panel. **b**, **c**, Higher-magnification images of the same cluster obtained in separate runs. Atomic-scale resolution is always restricted to tiny regions, containing about 30 unit cells (imaged at 283 K). Crosses in **b** indicate the positions (± 3 Å) where a subset of the spectra averaged in Fig. 3b was measured. **d**, **e**, Two other instances of tiny regions with atomic resolution at 283 K. The amplitude profiles show selected cross-sections from each micrograph ($Z(X,Y)$, blue lines), highlighting the half-unit-cell high steps (**a**) and the systematically enhanced tunnelling conductance (**b**) associated with atomic-scale resolution of the in-plane 3.86 Å unit-cell (**c**–**e**).

Spatially resolved tunnelling spectroscopy can establish electronic phase separation into metallic and insulating regions^{15,16} and determine whether there is a specific signature associated with the magnetic metal–insulator transition at $T_c = 90$ K. Therefore, for each temperature, we measured and analysed thousands of current–voltage characteristics, $I(V)$, at different locations over areas ranging from 10 to 10^6 nm². Several conclusions can be directly read from the representative data in Fig. 3a, b. First, the current distribution (white band) for each bias stemming from 2,000 spectra exactly matches the spread in each individual spectrum (dots), meaning that the spectra are extremely homogeneous over micrometre-sized regions at all temperatures. Furthermore, at 283 K, the regions with resolved atoms are spectroscopically indistinguishable (to within current experimental error) from surrounding featureless areas. Our most remarkable discovery is also directly visible in Fig. 3: the spectrum is gapped at 44 K (that is, essentially no current flows between two bias voltages, V_+ and V_-), seemingly at odds with the bulk transport measurements¹⁷ showing a two orders of magnitude drop in resistivity to metallic behaviour below 90 K.

To quantify the tunnelling characteristics, we evaluate the zero-bias conductance, σ_0 , namely the slopes of $I(V)$ at $V = 0$ (red line in Fig. 3a, b). For each temperature, the slopes from thousands of spectra fall on a normal distribution, gradually shifting away from zero on increasing the temperature (Fig. 3c, inset). The increase of spread with temperature is a purely instrumental effect, accounted for by a constant electronic noise enhanced by the temperature dependence of the STM's piezoelectric coefficients. Hence, at no temperature do our data show any sign of electronic phase separation into metallic and insulating regions, which would yield two peaks

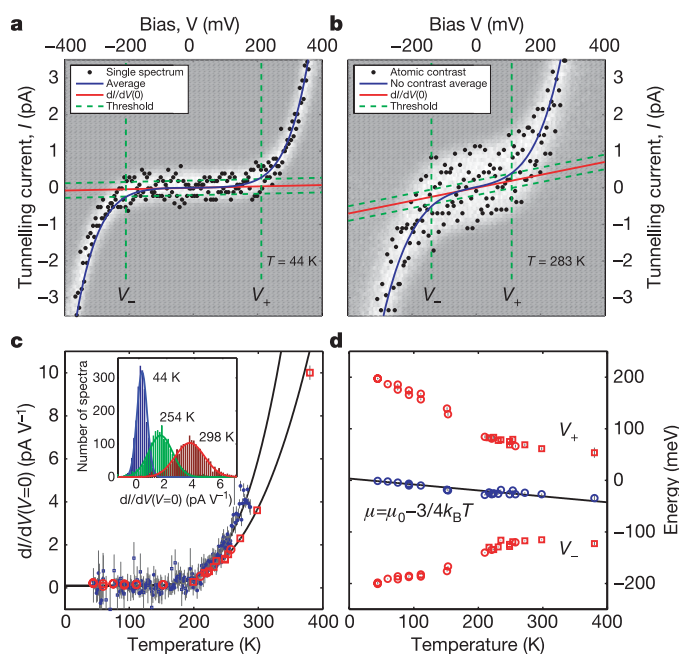


Figure 3 | Temperature-dependent STM spectroscopy of $\text{La}_{1.4}\text{Sr}_{1.6}\text{Mn}_2\text{O}_7$. **a**, **b**, $I(V)$ at 44 K (**a**) and at 283 K (**b**); a subset of sampling positions are indicated in Fig. 2b (see text for further details). **c**, Tunnelling conductance σ_0 at $V = 0$ as a function of temperature. The open symbols are obtained from the averages of thousands of $I(V)$ spectra measured over large areas at fixed temperatures. Filled symbols derive from single $I(V)$ characteristics measured on slowly warming. Fitting $\sigma_0(T)$ to $T^{3/2} \exp(-\Delta/2k_B T)$ gives $\Delta = 196 \pm 12$ meV for both sets of data. The inset shows the uniform distribution of σ_0 obtained from thousands of $I(V)$ characteristics measured at individual points on large areas at selected fixed temperatures. **d**, Temperature dependence of the gap edges V_- and V_+ , showing the persistence of the gap below the metal–insulator transition at 90 K. Blue circles indicate the mid-gap energy $(V_+ + V_-)/2$, which is independent of the chosen threshold (0.2 pA in this case) and corresponds to the chemical potential μ .

rather than the observed single peak in the distribution of σ_0 values.

The temperature (T) dependence of $\sigma_0(T) \propto T^{3/2} \exp(-\Delta/2k_B T)$, where k_B is Boltzmann's constant, is consistent with a thermally activated conductance across a gap $\Delta = 196 \pm 12$ meV at the Fermi energy (Fig. 3c), in agreement with the gap $\Delta_p = 188$ meV found from fitting the high-temperature c -axis resistivity (ρ_c) to the thermally activated form (Fig. 1b). Because the detailed nature of the tunnel junction changes between experiments, the absolute value of the zero-bias slope may vary (Fig. 3c). However, it does not affect the extracted gap, giving additional confidence that we are probing a gap intrinsic to LSMO.

Whereas σ_0 can be determined in a robust manner, it is more difficult to extract Δ from the tunnelling spectra. Bearing in mind that the zero-bias slope is simply due to thermal activation, we define V_- (V_+) as the negative (positive) bias at which the tunnelling current departs by a given amount (threshold) from the ohmic background (Fig. 3d). The gap amplitude, defined as $\Delta = V_+ - V_-$, depends on the particular choice of threshold, but its dramatic temperature dependence, doubling between room and base temperatures, much less so. Surprisingly, a more conventional linear behaviour is found for the chemical potential, $\mu = (V_+ + V_-)/2 \approx -3/4k_B T$, which, assuming the classic law $\mu = -3/4k_B T \ln(m_e/m_h)$ for an intrinsic semiconductor¹⁸, yields $m_e/m_h = 2.7$, implying that holes travel more easily along the c axis than electrons. (Here m_h (m_e) is the hole (electron) effective mass.)

Optical reflectivity is another probe of the electronic excitation spectra of solids, and has the advantages of much less surface sensitivity than photoemission and scanning tunnelling spectroscopy (STS), in addition to not depending on the details—crucial for d.c. conductivity—of electrical current flow in a macroscopic sample with electrical contacts. Consistency between optical data and STS is therefore a definitive demonstration of whether the STS measurements represent surface or intrinsic bulk phenomena. Accordingly, Fig. 4 shows the optical conductivity¹⁹ for LSMO. When polarized photons drive the carriers to move within the planes, the low-energy part of the spectra displays a marked Drude-like tail below T_c , signalling metallic behaviour (Fig. 4a) in excellent agreement with the in-plane resistivity data of Fig. 1b. In contrast, there is no sign of a metallic Drude-like peak in the c -axis optical conductivity (Fig. 4b), which shows a clear gap with no spectral weight up to about 400–500 meV (except for sharp phonon bands in the range 20–80 meV). The conductivity onset energy is entirely consistent with the low-temperature gap $\Delta = V_+ - V_-$ deduced from our STS spectra.

What is striking about our findings is the consistency of the STS and c -axis optical conductivity. Furthermore, the same gap value characterizes the temperature-dependent spectroscopy and the high temperature ρ_c , indicating that both measurements probe the same bulk property. However, on cooling we uncover the curious result that STS seems oblivious to the Néel point (90 K), below which ρ_c displays seemingly metallic behaviour while the zero-bias conductance, the STS spectra and $\sigma_c(\omega)$ are all consistent with an insulator (here ω is frequency). How can we reconcile the striking discrepancy below T_c between macroscopic resistivity measurements and the vacuum tunnelling and optical spectroscopies? The most obvious idea is to assert, as has been done on the basis of analysis of X-ray data for samples exposed to air, that the surface probed via tunnelling is somehow insulating and non-magnetic²⁰, and therefore different from the bulk sampled in electrical measurements. However, this is at odds with spin-polarized scanning electron microscopy clearly demonstrating that the top layer is ferromagnetic²¹, the in-plane metallic screening (see ref. 22 and discussion below) underlying the general lack of atomic contrast in ultrahigh-vacuum STM, static secondary ion mass spectroscopy suggesting that the top 1 nm of our cleaved sample has the same chemical composition as the bulk, and the agreement of STS and the much less surface-sensitive optical conductivity results.

The data show that whatever dominates the d.c. conductivity along c has no effect on the optical and STS measurements. To

account for the d.c. metallicity, we consider LSMO as a stack of alternating metallic and insulating layers. In a perfect sample, the c -axis transport would always be thermally activated, with ρ_c diverging at low temperature in agreement with the STM experiments. The sudden drop of ρ_c at the ferromagnetic transition must then be due to a parallel conduction mechanism. The idea is that while adjacent bilayers have predominantly antiparallel magnetizations, there will be defects where the domain walls in adjacent layers do not exactly coincide, resulting in regions containing adjacent bilayers whose magnetizations are parallel (Fig. 1a). In these regions the tunnelling probability will be enhanced, as demonstrated by the large and negative c -axis magnetoresistance seen in the low-temperature limit^{10,12}. Whereas the defective regions only contribute weakly to ρ_c above T_c , they turn into a well-connected network below T_c , effectively shunting the cation layers. Indeed, as depicted in Fig. 1b, $\rho_c(T)$ is reproduced by a simple model, described in the legend, with the activated intrinsic c -axis resistance $\rho_c^0(T) \propto T^{-3/2} \exp(\Delta_p/2k_B T)$ in parallel with the shunt resistance dominated by $\rho_{ab}(T)$ (where ρ_{ab} is the in-plane resistivity). As $\rho_c^0(T)$ diverges on lowering the temperature, the intrinsic anisotropy becomes larger and the c -axis resistance increasingly mimics the in-plane resistivity. In contrast, the parallel channel is inoperative in the STM tunnelling process, where a vacuum barrier is maintained between the tip and the sample surface, and so does not contribute to the temperature-dependent zero-bias conductance.

The key findings of our investigations are (1) position-independent and (2) gapped tunnelling and c -axis optical spectra at all temperatures, (3) metallic in-plane optical and d.c. conductivity, (4) c -axis d.c. conductivity that follows an activated form consistent with both STS and $\sigma_c(\omega)$ at high T but then tracks the in-plane d.c. conductivity rather than the STS zero-bias conductance at low T , and (5) atomic resolution that is stable and reproducible but limited to very small regions on the extremely flat surfaces prepared in ultrahigh

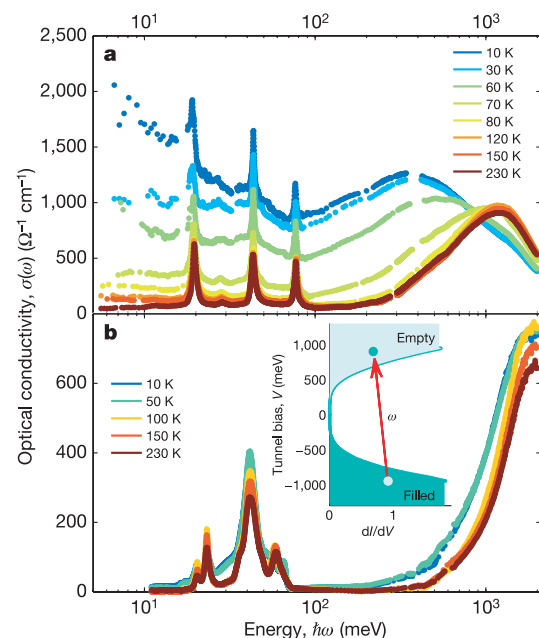


Figure 4 | Optical conductivity deduced from near-normal-incidence reflectivity measured on single crystal samples (sharp peaks at 20–80 meV are due to phonons)¹⁹. **a**, In-plane optical conductivity. The dramatic increase at low frequencies (ω) below 90 K corresponds to Drude-like metallic behaviour. **b**, The c -axis optical conductivity shows a semiconductor-like absorption edge at around 4–500 meV. Inset shows how the conductivity is related to a filled to empty state transition in the single particle density of states measured via STM (Fig. 3a, b). Imperfect alignment of the photon polarization and the crystal c -axis mixes the measured in-plane and out-of-plane conductivities, and is the most likely cause of the leftward shift of the absorption edge on cooling.

vacuum. All indications from the STM data themselves—including the step heights and lattice parameters measured when atomic resolution is found—demonstrate that the results are intrinsic to cleaved surfaces of LSMO. Furthermore, we have presented detailed evidence and analysis indicating that the STS, optical and d.c. conductivity measurements are entirely consistent with each other. This leaves us with the challenge of understanding the underlying anisotropy in quasiparticle transport out of the planes in LSMO and the surprising observation (5).

The quasiparticles in strongly correlated materials can be very different and more complex than the renormalized electrons found in simple metals. As STS extracts or inserts single (bare) electrons, the quasiparticles must dissociate in the tunnelling process. Even if the quasiparticles form a metal in the plane, STS will measure a gap corresponding to the energy it takes to destroy the quasiparticle to extract the electron (hole) from the MnO layer. Thus, when the transport, STS and optical data on LSMO are taken together, we are left with a picture where charge carriers are confined to metallic bilayers in LSMO, and cannot escape without a finite energy transfer. This is true down to the lowest temperatures, where our LSMO samples are magnetically ordered, and the conventional theory of solids would suggest a fully three-dimensional metal. The charge carriers are therefore bound to larger objects that can move within, but not between, the planes. Possible candidates for such objects are the polarons whose effects on the lattice and spin texture have been seen using X-rays and neutron scattering^{23,24}. Polarons have also been associated with the pseudogap observed by angle-resolved photoemission spectroscopy (ARPES)²⁵ in LSMO with $x = 0.4$ and the gap measured by STM²⁶ in the pseudo-cubic compound $\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$ with $x = 0$ and $x = 0.3$. Most recently, detailed ARPES experiments led to a very similar description of LSMO ($x = 0.4$) in terms of a polaronic metal with a gapped anisotropic electronic structure²⁷.

It turns out that the same polaron picture provides a similarly clear understanding of observation (5), and is at the heart of our discovery of atomic-scale features in a metallic manganite. Specifically, the general inability to resolve atoms—in contrast to the case for the charge-ordered manganite $(\text{Bi,Ca})\text{MnO}_3$ (ref. 28)—suggests metallic screening of charges within the a – b planes, in agreement with the bulk in-plane conductivity data. Typically, STM can only image metal atoms (for example, in gold) under conditions where there is bonding and relative motion between the tip and sample atoms. For the parameters achievable in our measurements, screening would reduce atomic contrast to below $0.02 \text{ \AA}$ (ref. 22). The small areas where atomic resolution is attained must therefore be regions where screening is relieved, owing perhaps to a point defect.

In some famous experiments²⁹ on much simpler metals, and recently on HTSs³⁰, quasiparticle scattering at defects results in electron density modulations with periodicities fixed by the quasiparticle dispersion relation. Obviously a sufficiently strong defect may bind (trap) a quasiparticle, resulting in a static (non-dispersive) STM pattern reflecting the internal structure of the polaron. There is some experimental evidence for this phenomenon in the copper oxides³¹ at energies below the quasiparticle dispersion effects. Here, we speculate that this is what is occurring when we see atoms in LSMO, and our experiments may have captured the real-space image of that very elusive object—a trapped polaron. The trapped polaron concept could account not only for the relative paucity of regions with resolved atoms on surfaces that were optimized for smoothness and cleanliness, but would also explain their common size, which in fact coincides with neutron and X-ray scattering results^{23,24} for the lattice and spin textures associated with polarons.

Received 6 September 2005; accepted 13 February 2006.

1. Anderson, P. W. The resonating valence bond state in La_2CuO_4 and superconductivity. *Science* **235**, 1196–1198 (1987).
2. Renner, Ch., Revaz, B., Genoud, J. Y., Kadowaki, K. & Fischer, Ø. Pseudogap

- precursor of the superconducting gap in under- and overdoped $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$. *Phys. Rev. Lett.* **80**, 149–152 (1998).
3. Uchida, S., Tamasaku, K. & Tajima, S. c-axis optical spectra and charge dynamics in $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$. *Phys. Rev. B* **53**, 14558–14574 (1996).
 4. von Helmolt, R., Wecker, J., Holzapfel, B., Schultz, L. & Samwer, K. Giant negative magnetoresistance in perovskitelike $\text{La}_{2/3}\text{Ba}_{1/3}\text{MnO}_x$ ferromagnetic films. *Phys. Rev. Lett.* **71**, 2331–2333 (1993).
 5. Jin, S. *et al.* Thousandfold change in resistivity in magnetoresistive La–Ca–Mn–O films. *Science* **264**, 413–415 (1994).
 6. Millis, A. J. Lattice effects in magnetoresistive manganese perovskites. *Nature* **392**, 147–150 (1998).
 7. Ramirez, A. P. *et al.* Thermodynamic and electron diffraction signatures of charge and spin ordering in $\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$. *Phys. Rev. Lett.* **76**, 3188–3191 (1996).
 8. Mori, S., Chen, C. H. & Cheong, S.-W. Pairing of charge-ordered stripes in $(\text{La,Ca})\text{MnO}_3$. *Nature* **392**, 473–476 (1998).
 9. Tokura, Y. & Nagaosa, N. Orbital physics in transition-metal oxides. *Science* **288**, 462–468 (2000).
 10. Perring, T. G., Aeppli, G., Kimura, T., Tokura, Y. & Adams, M. A. Ordered stack of spin valves in a layered magnetoresistive perovskite. *Phys. Rev. B* **58**, R14693–R14696 (1998).
 11. Dagotto, E. in *Nanoscale Phase Separation and Colossal Magnetoresistance* (eds Cardona, M. *et al.*) (Springer Series in Solid State Sciences Vol. 136, Springer, New York, 2003).
 12. Kimura, T. & Tokura, Y. Layered magnetic manganites. *Annu. Rev. Mater. Sci.* **30**, 451–474 (2000).
 13. Abrikosov, A. A. Quantum interference effects in quasi-two-dimensional metals. *Phys. Rev. B* **61**, 7770–7774 (2000).
 14. Moritomo, Y., Asamitsu, A., Kuwahara, H. & Tokura, Y. Giant magnetoresistance of manganese oxides with a layered perovskite structure. *Nature* **380**, 141–144 (1996).
 15. Becker, T. *et al.* Intrinsic inhomogeneities in manganite thin films investigated with scanning tunneling spectroscopy. *Phys. Rev. Lett.* **89**, 237203 (2002).
 16. Fäth, M. *et al.* Spatially inhomogeneous metal-insulator transition in doped manganites. *Science* **285**, 1540–1542 (1999).
 17. Kimura, T. *et al.* Interplane tunneling magnetoresistance in a layered manganite crystal. *Science* **274**, 1698–1701 (1996).
 18. Kittel, C. *Introduction to Solid State Physics* (Wiley and Sons, Inc., New York, 1976).
 19. Ishikawa, T., Tobe, K., Kimura, T., Katsufuji, T. & Tokura, Y. Optical study on the doping and temperature dependence of the anisotropic electronic structure in bilayered manganites: $\text{La}_{2-2x}\text{Sr}_{1+2x}\text{Mn}_2\text{O}_7$ ($0.3 \leq x \leq 0.5$). *Phys. Rev. B* **62**, 12354–12362 (2000).
 20. Freeland, J. W. *et al.* Full bulk spin polarization and intrinsic tunnel barriers at the surface of layered manganites. *Nature Mater.* **4**, 62–67 (2005).
 21. Konoto, M. *et al.* Direct imaging of temperature-dependent layered antiferromagnetism of a magnetic oxide. *Phys. Rev. Lett.* **93**, 107201 (2004).
 22. Blanco, J. M. *et al.* First-principles simulations of STM images: From tunneling to the contact regime. *Phys. Rev. B* **70**, 085405 (2004).
 23. Perring, T. G., Aeppli, G., Moritomo, Y. & Tokura, Y. Antiferromagnetic short-range order in a two dimensional manganite exhibiting giant magnetoresistance. *Phys. Rev. Lett.* **78**, 3197–3200 (1997).
 24. Campbell, B. J. *et al.* Structure of nanoscale polaron correlations in $\text{La}_{1.2}\text{Sr}_{1.8}\text{Mn}_2\text{O}_7$. *Phys. Rev. B* **65**, 014427 (2001).
 25. Dessau, D. S. *et al.* k -dependent electronic structure, a large “ghost” Fermi surface, and a pseudogap in a layered magnetoresistive oxide. *Phys. Rev. Lett.* **81**, 192–195 (1998).
 26. Wei, J. Y. T., Yeh, N. C. & Vasquez, R. P. Tunneling evidence of half-metallic ferromagnetism in $\text{La}_{0.7}\text{Ca}_{0.3}\text{MnO}_3$. *Phys. Rev. Lett.* **79**, 5150–5153 (1997).
 27. Mannella, N. *et al.* Nodal quasiparticle in pseudogapped colossal magnetoresistive manganites. *Nature* **438**, 474–478 (2005).
 28. Renner, Ch., Aeppli, G., Kim, B. G., Soh, Y. A. & Cheong, S.-W. Atomic-scale images of charge ordering in a mixed-valence manganite. *Nature* **416**, 518–521 (2002).
 29. Crommie, M. F., Lutz, C. P. & Eigler, D. M. Imaging standing waves in a two-dimensional electron gas. *Nature* **363**, 524–527 (1993).
 30. Hoffman, J. E. *et al.* Imaging quasiparticle interference in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$. *Science* **297**, 1148–1151 (2002).
 31. Yazdani, A., Howald, C. M., Lutz, C. P., Kapitulnik, A. & Eigler, D. M. Impurity-induced bound excitations on the surface of $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_8$. *Phys. Rev. Lett.* **83**, 176–179 (1999).

Acknowledgements We thank T. Ishikawa for providing the optical conductivity data shown in Fig. 4, D. McPhail and R. Chater for the static SIMS measurements, and A. Fisher and J. Mesot for discussions. We acknowledge support from a Wolfson–Royal Society Research Merit Award, the NEC corporation, and the European Commission through a FW6 STREP programme. H.M.R. thanks T. F. Rosenbaum for support through an NSF Materials Research Science and Engineering Centers grant.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Ch.R. (c.renner@ucl.ac.uk).

Climate sensitivity constrained by temperature reconstructions over the past seven centuries

Gabriele C. Hegerl¹, Thomas J. Crowley¹, William T. Hyde¹ & David J. Frame²

The magnitude and impact of future global warming depends on the sensitivity of the climate system to changes in greenhouse gas concentrations. The commonly accepted range for the equilibrium global mean temperature change in response to a doubling of the atmospheric carbon dioxide concentration¹, termed climate sensitivity, is 1.5–4.5 K (ref. 2). A number of observational studies^{3–10}, however, find a substantial probability of significantly higher sensitivities, yielding upper limits on climate sensitivity of 7.7 K to above 9 K (refs 3–8). Here we demonstrate that such observational estimates of climate sensitivity can be tightened if reconstructions of Northern Hemisphere temperature over the past several centuries are considered. We use large-ensemble energy balance modelling and simulate the temperature response to past solar, volcanic and greenhouse gas forcing to determine which climate sensitivities yield simulations that are in agreement with proxy reconstructions. After accounting for the uncertainty in reconstructions and estimates of past external forcing, we find an independent estimate of climate sensitivity that is very similar to those from instrumental data. If the latter are combined with the result from all proxy reconstructions, then the 5–95 per cent range shrinks to 1.5–6.2 K, thus substantially reducing the probability of very high climate sensitivity.

We use four palaeoreconstructions, namely a hemispheric reconstruction of mean annual temperatures¹¹, a maximum latewood density tree ring based reconstruction¹² for growing season temperatures over 20–90° N land, a revised and smoothed version of a record¹³ that has been calibrated to 30–90° N land annual data¹⁴, and our own new decadal reconstruction termed ‘CH-blend’ of annual average 30–90° N temperature¹⁵ (Fig. 1). A version of CH-blend using 12 records extends from AD 1505 to AD 1960; and a reconstruction based on 9 sites (‘CH-blend (long)’) is used from AD 1270. Both reconstructions use a relatively small number of well spaced sites (often based on multiple records, including some regional reconstructions) throughout the reconstruction. CH-blend is consistent with independent estimates of temperatures from boreholes¹⁵, and both CH-blend and CH-blend (long) agree well with a recent reconstruction¹⁶ that incorporates records of lower temporal resolution. The reconstruction method has been tested using noise-perturbed climate model data from the same locations as used in the reconstruction¹⁵. Results show that the reconstruction of decadal temperatures is accurate and reliably preserves the variance of hemispheric-scale temperature variability.

For CH-blend, our estimate of climate sensitivity fully accounts for the uncertainty in the amplitude of the record¹⁵. For the other reconstructions, we use both the published reconstruction and a version that is recalibrated using our technique. This approach avoids introducing a low bias in our estimate of climate sensitivity based on the possibility that some reconstruction

techniques underestimate past climate variability¹⁷ (for details see Supplementary Information).

We conduct a large ensemble (>1,000) of simulations of the past 1,000 years with a 2.5-dimensional (latitude/longitude/depth) upwelling-diffusion energy balance model (EBM), with realistic land-sea distribution. The EBM is a variant of a seasonal model¹⁸ that simulates time-dependent responses to external forcing, and includes the seasonal cycle (for details see Supplementary Information). The same model has been previously used to examine the relationship between reconstructed temperature and external forcing over the past millennium^{19,20}. EBM simulations reproduce the large-scale temperature response of general circulation models, and have the advantage of being able to generate large ensembles.

The following two model parameters are important determinants of the large-scale response of climate models to external forcing⁵ and have been systematically varied in our ensemble. First, the equilibrium climate sensitivity to a doubling of CO₂, α , which was varied in 0.5 K increments from 0.5 K to 10.0 K with an additional low value of 0.25 K. Second, effective ocean diffusivity κ in the upwelling-diffusive model²¹, which was varied between 0.63 cm² s⁻¹ and 3.8 cm² s⁻¹.

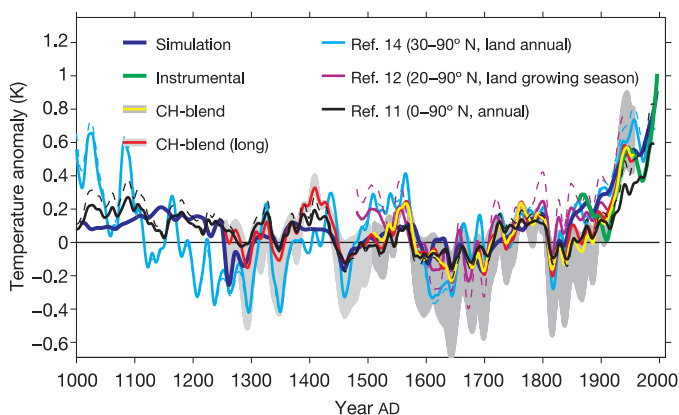


Figure 1 | Palaeoclimatic records compared to a climate model simulation. ‘CH-blend’ and ‘CH-blend (long)’ represent 30–90° N annual mean temperature (grey shading shows 10–90% ranges for uncertainty in the amplitude of the reconstruction); ref. 11 shows 0–90° N land temperature; ref. 14 shows 30–90° N land temperature; and ref. 12 shows 20–90° N land growing season temperature (dashed line indicates reconstructions rescaled¹⁵). The model (‘Simulation’) has a sensitivity of 2.5 K, mid-range ocean diffusivity and is driven with mid-range aerosol forcing. All data are smoothed to focus on multi-decadal variability and shown as anomalies relative to the period before 1800. The instrumental record for 30–90° N annual mean surface temperature (‘Instrumental’) is offset to match CH-blend between 1880 and 1960.

¹Division of Earth and Ocean Sciences, Nicholas School of the Environment and Earth Sciences, Duke University, Durham, North Carolina, 27708, USA. ²Climate Dynamics Group, Department of Physics, University of Oxford, Oxford OX1 3PU, UK.

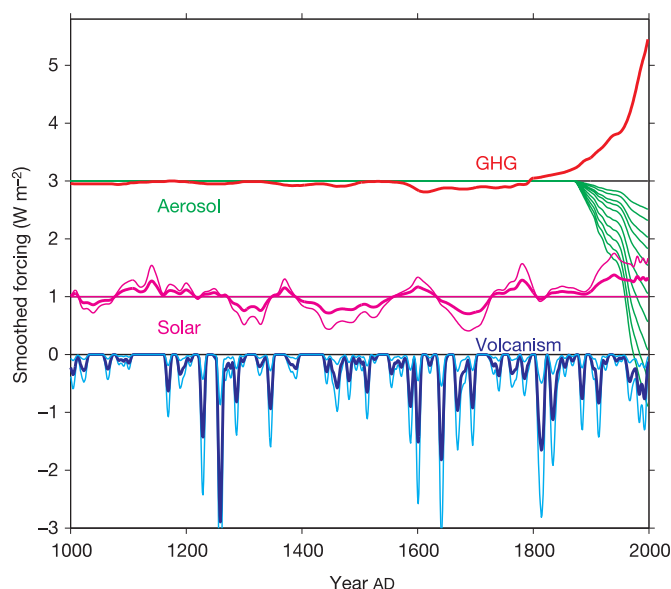


Figure 2 | Northern Hemisphere mean radiative forcing. Sub-annual forcing data are used in the climate model simulations, but a decadal filter is applied here for illustration only, to focus on timescales most relevant for the analysis. For tropospheric aerosol forcing (green), a range of forcing has been used; for solar (pink) and volcanic (blue) forcing, a best guess forcing (dark, thick line) and a gaussian uncertainty range has been used (2.5% and 97.5% limits are shown by light, thin lines, the lower limit for solar is on the zero line). For clarity, greenhouse gas ('GHG') and aerosol forcings are offset by 3 W m^{-2} , and solar forcing by 1 W m^{-2} .

This range embraces an observational estimate of $1.7 \pm 0.2 \text{ cm}^2 \text{ s}^{-1}$ based on a global compilation of GEOSECS data of bomb tritium penetration into the world ocean²² and a lower range²³ based on bomb ^{14}C of the order of $1 \text{ cm}^2 \text{ s}^{-1}$. We have further tested our range of diffusivities by comparing simulated ocean warming with ocean heat content data²⁴. We find that the smaller to mid-range values of κ yield results that compare most favourably with these data, consistent with the observation that most of the twentieth-century increase in ocean heat content is in the upper 1,000 m (Supplementary Fig. 1). Note that ocean diffusivity is of smaller importance for the simulations of the pre-industrial period, where forcings are mostly episodic and relatively small, than for the twentieth century. In the latter period, the rate of temperature increase is crucially influenced by ocean diffusivity, as large diffusivities tend to hide more warming

in the oceans than small diffusivities (see Supplementary Information for more discussion). Our results are insensitive to attempts to constrain κ further. They are, however, conditional on ocean effective diffusivity being within the range we use.

Prior work^{15,19,20,25} has established that various reconstructions of hemispheric temperature consistently show influence from volcanism and greenhouse gas variations, and less consistently from variations in solar radiation. We force the EBM simulations with a combination of solar, volcanic, greenhouse gas and tropospheric aerosol forcing to simulate hemispheric temperature change over the past millennium (Fig. 2). Greenhouse gas forcing is based on changes in trace gases from ice-core data, combined with IPCC estimates of radiative forcing for well-mixed greenhouse gases in the twentieth century. The estimate of solar forcing is based on ^{14}C data²⁶, scaled to the solar irradiance reconstruction of ref. 27 after reducing its amplitude by 20% to accommodate recent conclusions that the former estimate may have been large²⁸. For volcanism, we use an update of a global reconstruction²⁰ based on ice-core data from Greenland and Antarctica. We account for the considerable uncertainty in solar and volcanic forcing by varying the total amplitude of each forcing time-series around its central estimate. We use Monte Carlo simulations based on a 50% standard deviation for solar forcing, and a 35% standard deviation for volcanic forcing (excluding the unphysical case of net negative forcing). The uncertainty in our results due to random errors in the magnitude of individual volcanic eruptions was estimated by sensitivity tests, indicating that errors in the magnitude of individual eruptions can cause a modest widening of the tail of the distribution (Supplementary Fig. 2; see Supplementary Information for more detail on forcings and their uncertainty).

We derive a probability density function (PDF) for climate sensitivity using a method related to one previously used for instrumental data^{5,6} (see Methods section, and algorithm in Supplementary Information). Results for the CH-blend reconstruction, for which we have the most reliable uncertainty estimate¹⁵, yield a 5–95% range for sensitivity of 1.4 K to 6.1 K and a median sensitivity of 2.6 K over the pre-instrumental period 1505–1850 (Fig. 3a). PDFs for climate sensitivity from the other reconstructions and the same period yield peak probabilities (modes) from 1.3 K to 3.6 K, and some of them suggest a moderate probability for climate sensitivity being high (see Supplementary Table 4). Reconstructions with higher amplitudes of past climate fluctuations generally suggest higher climate sensitivities than those with low variability. The range of the other free parameters, ocean diffusivity κ , and solar and volcanic forcing uncertainty, are used to fully explore uncertainties rather than to provide posterior information about best-fit values.

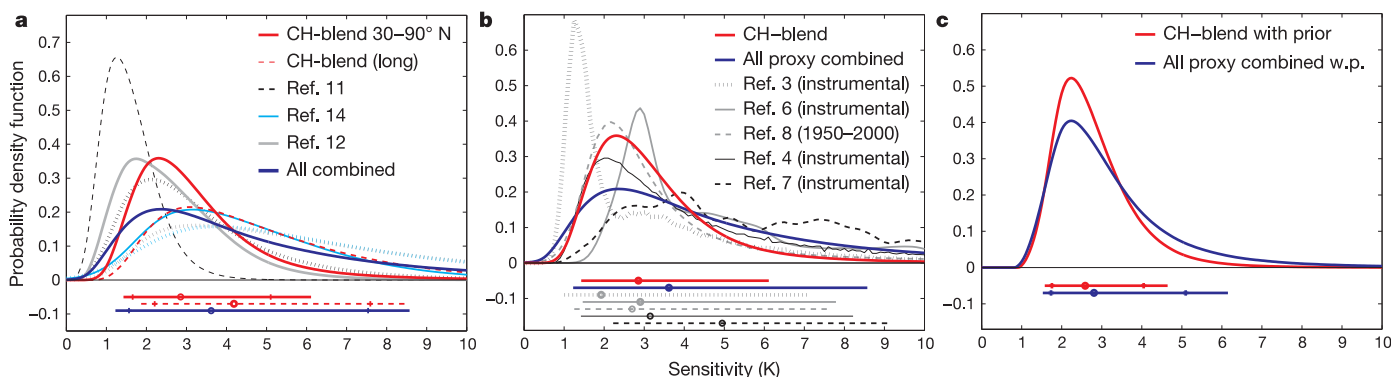


Figure 3 | Estimated probability density functions (PDFs) for equilibrium climate sensitivity to CO_2 doubling (in K). **a**, PDFs from a range of palaeoreconstructions using data to 1850 (dotted lines, based on rescaled data). The horizontal bars indicate the 5–95% range of PDFs (median is indicated by a dot, and 10th and 90th percentiles by a vertical bar).

b, Comparison to other estimates of climate sensitivity based on instrumental data^{3,4,6,7} over the twentieth century or 1950–2000⁸. All PDFs have been scaled to integrate to 1 between 0 and 10 for better comparison. **c**, Combined estimate using a result from instrumental data⁸ as prior distribution, updated by the result from pre-industrial data ('w.p.').

Consistent with that, the pre-industrial period does not provide constraints for ocean diffusivity, nor do results consistently favour a particular realization of forcing uncertainty, apart from a general preference for solar forcing on the low end of the range.

If all four reconstructions, both published and rescaled, are considered as equally likely realizations of the true hemispheric temperature evolution, the PDF that describes results from all four reconstructions combined yields a median sensitivity of 3.4 K and a 5–95% range of 1.2 K to 8.6 K. This renders negative climate feedbacks to CO₂ changes (corresponding to a climate sensitivity of under 1.1 K) very unlikely. As in other estimates of climate sensitivity, the upper tail is not well constrained. Note that the upper limit of the transient climate response, which governs the near-term magnitude of the climate response, tends to be better constrained from observations than equilibrium climate sensitivity⁸.

Our results are remarkably consistent with PDFs for climate sensitivity that have been estimated from the instrumental record^{13–8} (Fig. 3b) and that account for a differing level of uncertainty in forcings (most notably aerosol forcing⁷), ocean diffusivity and observations. The response of climate to pre-industrial forcing is governed (to a very reasonable approximation) by the same climate sensitivity. However, the uncertainties affecting each estimate are virtually independent, as the pre-industrial temperature reconstructions are virtually independent from instrumental temperatures for the second half of the twentieth century (decadal data before that are used for calibrating the palaeodata) and different forcing uncertainties affect each estimate. Therefore we can combine results from both to further constrain sensitivity. We use a version of the joint PDF for diffusivity and sensitivity, κ and α , from ref. 8 that is based on decadal instrumental data from 1950 to 2000 as a prior probability distribution (the use of a prior PDF of κ and α combined accounts for their dependence; Supplementary Fig. 5 shows a comparison between the published results for the entire twentieth century⁸ and the prior used here). We have widened the upper tail of the ref. 8 estimate in sensitivity in order to conservatively account for further uncertainties and embrace other instrumental estimates (results are only moderately sensitive to this, for details see Supplementary Information). Bayes' theorem is then used to calculate a posterior probability based on data from the past millennium (Fig. 3c). The resulting 5–95% ranges for CH-blend shrink to 1.6 K to 4.6 K, and those for all proxy data combined to 1.5 K to 6.2 K. This result reduces the probability from 36% to 15% or less that climate sensitivity exceeds the upper limit of the IPCC range of 4.5 K.

As previously shown, the agreement between models and data is mostly driven by the temperature response to volcanism, which causes longer-term variability due to the changing statistics of volcanic eruptions (Fig. 2; see also Supplementary Table 3 for correlations between simulations and records on short and long timescales). A superposed epoch analysis previously showed that the EBM simulates the response-characteristics to volcanism very well¹⁹. The EBM does not simulate changes in atmospheric dynamics that have been associated with strong volcanic eruptions²⁹, but these changes do not much affect hemispheric annually (or growing season) averaged temperatures³⁰.

We also note that model uncertainties (beyond those that we account for) potentially affect all estimates of climate sensitivity. Although our results are conditional on the range of effective ocean diffusivity and the upwelling parameter being realistic, simulated ocean heat content changes in our best-fit simulation compare very well with recent data²⁴ (Supplementary Fig. 1). A simulation with the most likely sensitivity of 2.5 K also compares well to the low-frequency component of annual global temperatures from instrumental data (Supplementary Fig. 4).

We conclude that proxy-reconstructions of the pre-industrial period from 1270 to 1850 yield very similar estimates of climate sensitivity to those obtained from the virtually independent climate change over the twentieth century. This agreement increases our

confidence in the overall reliability of these estimates based on twentieth-century changes. When both independent lines of evidence are combined, the resulting PDF for climate sensitivity narrows, yielding a very small probability for climate sensitivity exceeding 7 K (<3% based on all reconstructions combined, and <1% based on CH-blend).

METHODS

Estimating climate sensitivity. Our method of estimating the PDF of equilibrium climate sensitivity is related to a method used previously for instrumental data^{5,6} and is briefly discussed here. A detailed algorithm can be found in the Supplementary Information. We simulate the time-space evolution of surface temperature over the past millennium forced with observed changes in solar, volcanic, greenhouse gas and sulphate aerosol forcing. We use a very large ensemble of EBM simulations with varying climate sensitivity α and ocean diffusivity κ , forced by different realizations of solar and volcanic forcing (f_{sol} , f_{vol}) to account for the most important uncertain parameters driving the simulated response. For the twentieth century, aerosol forcing, f_{aer} , is also varied. Each model simulation yields a time-space pattern of surface temperature response for each parameter $T(x, t, \alpha, \kappa, f_{\text{sol}}, f_{\text{vol}}, f_{\text{aer}})$ that depends on time t and space x . For each palaeoreconstruction $\bar{T}_{\text{palaeo}}$ (overbar denoting a spatial average), the simulated spatial temperature patterns are averaged over the latitude strip and season the reconstruction is calibrated to, and filtered to the time-resolution used for the analysis (annual reconstructions are filtered by a 5-yr running mean). All data are then centred from the beginning of the record to 1800 to focus on deviations from a mean climatic state of the past millennium and to avoid major variations in the climate state caused by anthropogenic forcing. We analyse reconstructions from the beginning of the record, but not before AD 1270 (as there are significant uncertainties in the radiative forcing effects of a very large eruption in 1258). The analysis focuses on the residual between the simulated record $\bar{T}$ and the observed record:

$$\text{res}(t, \alpha, \kappa, f_{\text{sol}}, f_{\text{vol}}, f_{\text{aer}}) = \bar{T}_{\text{palaeo}}(t) - \bar{T}(t, \alpha, \kappa, f_{\text{sol}}, f_{\text{vol}}, f_{\text{aer}}) \quad (1)$$

Some combination of parameters will yield the residual with the smallest estimated variance, $\hat{\sigma}_{\text{min}}^2 = \min_{\text{parameters}} \left(\frac{\|\text{res}(t, \alpha, \kappa, f_{\text{sol}}, f_{\text{vol}}, f_{\text{aer}})\|^2}{n-1} \right)$, with $\|\text{res}\|^2 = \sum \text{res}(t, \alpha, \kappa, f_{\text{sol}}, f_{\text{vol}}, f_{\text{aer}})^2$ and n denoting the length of the residual time series (yr). The difference between any square residual $\|\text{res}\|^2$ and the minimum square residual $\|\text{res}_{\text{min}}\|^2$ will then be F-distributed⁵

$$\frac{\|\text{res}(\alpha, \kappa, f_{\text{aer}}, f_{\text{sol}}, f_{\text{vol}})\|^2 - \|\text{res}_{\text{min}}\|^2}{\hat{\sigma}_{\text{min}}^2} \propto mF(m, \nu) \quad (2)$$

where m is the number of free parameters (4 for the pre-industrial case, 5 for the entire time-series) and ν is the number of degrees of freedom in res_{min} . Note that for autocorrelated data, ν will be smaller than n . Therefore we account for the number of effectively independent samples in the square residual (see Supplementary Information). Thus, the likelihood of the reconstruction given each set of model parameters and forcings $p(\text{data}|\alpha, \kappa, f_{\text{aer}}, f_{\text{vol}}, f_{\text{sol}})$ is estimated from the probability that its residual variability is statistically indistinguishable from the best-fit residual⁵, given internal climate variability and non-climatic random errors in proxy data.

Bayes' theorem is used to derive the joint PDF of the parameters

$$p(\alpha, \kappa, f_{\text{aer}}, f_{\text{vol}}, f_{\text{sol}}|\text{data}) \propto p(\text{data}|\alpha, \kappa, f_{\text{aer}}, f_{\text{vol}}, f_{\text{sol}}) \cdot p(\alpha, \kappa, f_{\text{aer}}, f_{\text{vol}}, f_{\text{sol}}) \quad (3)$$

from the likelihood of the data and the prior probability of the parameters $p(\alpha, \kappa, f_{\text{aer}}, f_{\text{vol}}, f_{\text{sol}})$. For results based on the proxy data alone, a uniform prior distribution for α is used for integration, which extends to a sensitivity of 10 K and then drops off to zero. Similarly, a prior for κ is used that is uniform over the range we cover (see main text), and normal prior distributions are used for solar and volcanic uncertainty (see Fig. 2). Where a prior distribution from the late twentieth century is used, the joint probability density function $p(\alpha, \kappa)$ from ref. 8 is applied instead of the uniform distribution.

This analysis is performed for every reconstruction. For reconstructions where the amplitude uncertainty can be fully accounted for (CH-blend and CH-blend (long)), this analysis is performed for the reconstruction scaled by a range of scaling factors β , representing uncertainty in the amplitude of the reconstruction (see Supplementary Information). We use the best-guess scaling and the 2.5th, 10th, 25th, 75th, 90th and 97.5th percentile of β . All PDFs resulting from these analyses $p(\text{data}(\beta)|\alpha, \kappa, f_{\text{aer}}, f_{\text{vol}}, f_{\text{sol}})$ are then averaged over β , weighted by the likelihood of each scaling β based on a normal distribution. This is a robust way of incorporating uncertainty in the reconstruction, as random errors in the reconstruction are directly accounted for in the residual.

For reconstructions where the amplitude uncertainty cannot be fully estimated, we use both the published best guess and a rescaled best guess using our calibration method to estimate $p(\text{data}|\alpha, \kappa, f_{\text{aer}}, f_{\text{vol}}, f_{\text{sol}})$.

The resulting multi-dimensional likelihood is integrated over κ and forcing uncertainties $f_{\text{sol}}, f_{\text{vol}}$ and f_{aer} , yielding a PDF for climate sensitivity. This is done both for each reconstruction individually, and for the average of the joint probabilities $p(\alpha, \kappa, f_{\text{aer}}, f_{\text{vol}}, f_{\text{sol}}|\text{data})$ from all reconstructions to derive an estimate of climate sensitivity from all records combined.

Our method to estimate sensitivity has been validated using synthetic data (see Supplementary Fig. 2), and tested by using an alternative method to estimate the likelihood based on scaling factors (see Supplementary Information).

Received 8 July 2005; accepted 28 February 2006.

- Schneider, S. H. & Dickinson, R. E. Climate modeling. *Rev. Geophys. Space Phys.* **12**, 447–493 (1974).
- Cubasch, U., et al. in *Climate Change 2001: The Scientific Basis* (eds Houghton, J. T. et al.) 525–582 (Cambridge Univ. Press, New York, 2001).
- Andronova, N. G. & Schlesinger, M. E. Causes of global temperature changes during the 19th and 20th centuries. *Geophys. Res. Lett.* **27**, 2137–2140 (2000).
- Gregory, J. M., Stouffer, R. J., Raper, S. C., Stott, P. A. & Rayner, N. A. An observationally based estimate of the climate sensitivity. *J. Clim.* **15**, 3117–3121 (2002).
- Forest, C. E., Stone, P. H., Sokolov, A. P., Allen, M. R. & Webster, M. D. Quantifying uncertainties in climate system properties with the use of recent observations. *Science* **295**, 113–117 (2002).
- Forest, C. E., Stone, P. H. & Sokolov, A. P. Estimated PDFs of climate system properties including natural and anthropogenic forcings. *Geophys. Res. Lett.* (in the press).
- Knutti, R., Stocker, T. F., Joos, F. & Plattner, G.-K. Constraints on radiative forcing and future climate change from observations and climate model ensembles. *Nature* **416**, 719–723 (2002).
- Frame, D. J. et al. Constraining climate forecasts: The role of prior assumptions. *Geophys. Res. Lett.* **32**, L09702, doi:10.1029/2004GL0022241 (2005).
- Murphy, J. M. et al. Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature* **430**, 768–772 (2004).
- Stainforth, D. A. et al. Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature* **433**, 403–406 (2005).
- Mann, M. E. & Jones, P. D. Global surface temperatures over the past two millennia. *Geophys. Res. Lett.* **30**, 1820, doi:10.1029/2003GL017814 (2003).
- Brieffa, K. R. et al. Low-frequency temperature variations from a northern tree ring density network. *J. Geophys. Res.* **106**, 2929–2941 (2001).
- Esper, J., Cook, E. R. & Schweingruber, F. Low-frequency signals in long tree-ring chronologies for reconstructing past temperature variability. *Science* **295**, 2250–2253 (2002).
- Cook, E., Esper, J. & D'Arrigo, R. Extra-tropical Northern Hemisphere land temperature variability over the past 1000 years. *Quat. Sci. Rev.* **23**, 2063–2074 (2004).
- Hegerl, G. C. et al. Detection of human influence on a new, validated 1500 yr temperature reconstruction. *J. Clim.* (submitted).
- Moberg, A., Sonechkin, D. M., Holmgren, K., Datsenko, N. M. & Karlen, W. Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy data. *Nature* **433**, 613–617 (2005).
- von Storch, H. et al. Reconstructing past climate from noisy data. *Science* **306**, 679–682 (2004).
- North, G. R., Mengel, J. G. & Short, D. A simple energy balance model resolving the seasons and continents: Application to the astronomical theory of the ice ages. *J. Geophys. Res.* **88**, 6576–6586 (1983).
- Hegerl, G. C., Crowley, T. J., Baum, S. K., Kim, K.-Y. & Hyde, W. T. Detection of volcanic, solar, and greenhouse gas signals in paleo-reconstructions of Northern Hemispheric temperature. *Geophys. Res. Lett.* **30**, 1242, doi:10.1029/2002GL016635 (2003).
- Crowley, T. J., Baum, S. K., Kim, K.-Y., Hegerl, G. C. & Hyde, W. T. Modeling ocean heat content changes during the last millennium. *Geophys. Res. Lett.* **30**, 1932, doi:10.1029/2003GL017801 (2003).
- Crowley, T. J. & Kim, K.-Y. Modeling the temperature response to forced climate change over the last six centuries. *Geophys. Res. Lett.* **26**, 1901–1904 (1999).
- Li, Y.-H., Peng, T.-H., Broecker, W. S. & Oestlund, H. G. The average vertical mixing coefficient for the oceanic thermocline. *Tellus B* **36**, 212–217 (1984).
- Siegenthaler, U. Uptake of excess CO₂ by an outcrop-diffusion model of the ocean. *J. Geophys. Res.* **88**, 3599–3608 (1983).
- Levitus, S., Antonov, J. & Boyer, T. Warming of the world ocean, 1955–2003. *Geophys. Res. Lett.* **32**, L02604, doi:10.1029/2004GL021592 (2005).
- Free, M. & Robock, A. Global warming in the context of the little Ice Age. *J. Geophys. Res.* **104**, 19057–19070 (1999).
- Stuiver, M. & Braziunas, T. F. Modeling atmospheric ¹⁴C influence and ¹⁴C ages of marine samples to 10,000 BC. *Radiocarbon* **35**, 137–189 (1993).
- Lean, J. L., Beer, J. & Bradley, R. Reconstruction of solar irradiance changes since 1610: Implications for climate change. *Geophys. Res. Lett.* **22**, 3195–3198 (1995).
- Lean, J. L., Wang, Y.-M. & Sheeley, N. R. The effect of increasing solar activity on the Sun's total open magnetic flux during multiple cycles: Implications for solar forcing of climate. *Geophys. Res. Lett.* **29**, 2224, doi:10.1029/2002GL015880 (2002).
- Shindell, D. T., Schmidt, G. A., Mann, M. E. & Faluvegi, G. Dynamic winter climate response to large tropical volcanic eruptions since 1600. *J. Geophys. Res.* **109**, D05104, doi:10.1029/2003JD004151 (2004).
- Thompson, D., Wallace, J. M. & Hegerl, G. C. Annular modes in the extratropical circulation: Part II, Trends. *J. Clim.* **13**, 1018–1036 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Forest for suggesting merging PDF estimates from results based on past-millennium climate changes with those based on twentieth-century climate change; J. Kenyon for technical support; M. Allen, F. Zwiers, A. Gelfand and M. Lavine for discussions; and C. Forest, N. Andronova, R. Knutti and J. Gregory for providing data. G.C.H. and T.J.C. were supported by NOAA, NOAA's office of global programmes and the DOE's office of biological and environmental research. G.C.H. was additionally supported by NSF.

Author Contributions G.C.H. developed and implemented the method to estimate sensitivity and to calibrate proxy records, T.J.C. provided the reconstruction of past forcing and developed the CH-blend reconstruction, W.T.H. performed the EBM simulation, and D.J.F. derived the prior estimate of sensitivity from instrumental data 1950–2000. G.C.H., T.J.C. and W.T.H. analysed the results.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.C.H. (hegerl@duke.edu).

Rapid discharge connects Antarctic subglacial lakes

Duncan J. Wingham¹, Martin J. Siegert², Andrew Shepherd^{3†} & Alan S. Muir¹

The existence of many subglacial lakes¹ provides clear evidence for the widespread presence of water beneath the East Antarctic ice sheet, but the hydrology beneath this ice mass is poorly understood². Such knowledge is critical to understanding ice flow, basal water transfer to the ice margin, glacial landform development and subglacial lake habitats. Here we present ice-sheet surface elevation changes in central East Antarctica that we interpret to represent rapid discharge from a subglacial lake. Our observations indicate that during a period of 16 months, 1.8 km³ of water was transferred over 290 km to at least two other subglacial lakes. While viscous deformation of the ice roof above may moderate discharge, the intrinsic instability of such a system³ suggests that discharge events are a common mode of basal drainage⁴. If large lakes, such as Lake Vostok or Lake Concordia¹, are pressurizing, it is possible that substantial discharges could reach the coast^{5,6}. Our observations conflict with expectations that subglacial lakes have long residence times and slow circulations^{2,7,8}, and we suggest that entire subglacial drainage basins may be flushed periodically. The rapid transfer of water between lakes would result in large-scale solute and microbe relocation, and drainage system contamination from *in situ* exploration is, therefore, a distinct risk.

The Adventure subglacial trench lies beneath Dome C in central East Antarctica (Fig. 1a). Ice thickness within the trench is over 4 km in places, which allows the bulk of the ice base to be at the pressure melting point⁹. A series of subglacial lakes has been identified along the axis of the trench from radio-echo sounding¹. Although the trench rises to the south, the overlying ice thickness decreases and the net effect is to cause the hydraulic potential¹⁰ to fall southwards. Subglacial water will flow south along the trough axis to the Wilkes subglacial basin, from where it may run to the ice-sheet margin (Supplementary Figs 1 and 2).

The altimeter survey¹¹ by the satellite ERS-2, at orbit-crossing points of this sector of the East Antarctic Ice Sheet from 1995 to 2003, reveals two clustered anomalies of ice-sheet surface elevation change in the vicinity of the Adventure subglacial trench (Fig. 1a). Close examination of one anomaly, at the northern end of the trench, reveals an abrupt fall in ice-surface elevation at three neighbouring crossover points ('L' sites) in 1997 (for example, curve L1 in Fig. 2). Some 290 km distant from L1, a corresponding abrupt rise occurred at three crossover points ('U' sites) that lie close to one another at the southern end of the trench (Fig. 1a, and U1, U2 and U3 in Fig. 2). The only mechanism that explains these observations is a rapid transfer of basal water from a subglacial lake beneath the region of surface lowering to lakes beneath the regions of uplift (Supplementary Fig. 1)¹². That lakes exist at the 'U' sites is known independently from radio-echo sounding records, and while no such records exist for the 'L' sites, they fall close to the flow-path estimated from the hydrological potential¹¹ upstream of the 'U' sites (Fig. 1a).

To determine the magnitude and rate of the subglacial discharge,

an estimate of the lake area at the 'L' sites is required. We attempted, using ERS-2 satellite radar interferometry¹³, to determine the spatial pattern of vertical ice displacement, from which the lake area can be measured. Unfortunately only one pair of images, capturing the final stage of the fall and lying north of the 'L' sites, provided a coherent signal (Fig. 1b). Nonetheless, the data reveal the northern end of a sharply defined oval pattern of subsidence whose axis coincides with the line of 'L' sites. It appears that the 'L' sites sample the lowering water level over a single subglacial lake (referred to hereafter as 'lake L'). Making an approximate (see Supplementary Methods 1) extrapolation to obtain its area (600 km²) and average fall (3 m), we estimate a total discharge of 1.8 km³. We also estimate from Fig. 2 that the peak discharge rate is approximately 50 m³ s⁻¹ (to give a practical idea of the scale, this is about three-quarters of the modal discharge of the Thames river in London, UK).

Rapid discharges through tunnels are a well-known phenomenon of temperate ice masses (for example, jökulhlaups)¹⁴. Their essential mechanism is well-established³: potential energy released by the flow melts the ice walls of the tunnel, and a positive feedback causes the flow to rise abruptly. The same mechanism, involving a single tunnel draining the lake, explains our observations (Supplementary Methods 2). With an average hydraulic gradient of 5.1 Pa m⁻¹, a discharge of 50 m³ s⁻¹ may be supported by a single semicircular tunnel of 4 m in radius with a Manning roughness coefficient of 0.08 m^{-1/3} s (refs 15, 16), and a 290-km tunnel of this radius is in keeping with the 2.7 × 10¹⁵ J of energy (Supplementary Methods 2) available to melt its volume. The termination of a discharge is usually explained as the result of tunnel closure by ice flow as the reservoir pressure falls. In the present case, however, the total change in reservoir pressure is only 29 kPa. While the effective pressure need not equal zero for a flood to initiate⁴, an effective pressure of 590 kPa would be required for the closure rate to exceed the melt rate of a single channel. A more obvious explanation for the termination is that lake L either emptied, or that the bed, newly connecting with the roof as the lake level fell, caused a new lake seal to form.

The data are not sufficient to restrict the drainage to a single (or a few) tunnels. With the single-channel average velocity of 2.1 m s⁻¹, the water would take 1.6 days to transit between the sites, but equally, velocities as low as 0.1 m s⁻¹, allowing numerous channels to be involved, would also remain difficult to resolve in the data of Fig. 2. Nor are the data sufficient to identify the detailed nature of the water pathway; it may be, for example, that eroded channels exist in the bed⁵ and it is their ceilings that are subject to melting. It is also possible that a number of smaller lakes or ponds along the way retain water following the termination of the discharge. Although it seems likely from Fig. 2 that the 'U' lakes were the destination of the bulk of the discharge, we are unable to account with useful accuracy for all the discharge mass. We found no evidence of significant water storage elsewhere in the Adventure subglacial trench, so any intermediate storage must be small or widely distributed. We consider it

¹Centre for Polar Observation and Modelling, Department of Space and Climate Physics, Pearson Building, University College London, Gower Street, London WC1E 6BT, UK.

²Centre for Polar Observation and Modelling, Bristol Glaciology Centre, School of Geographical Sciences, University of Bristol, Bristol, BS8 1SS, UK. ³Centre for Polar Observation and Modelling, Scott Polar Research Institute, University of Cambridge, Lensfield Road, Cambridge CB2 1ER, UK. †Present address: School of Geosciences, University of Edinburgh, Drummond Street, Edinburgh EH8 9XP, UK.

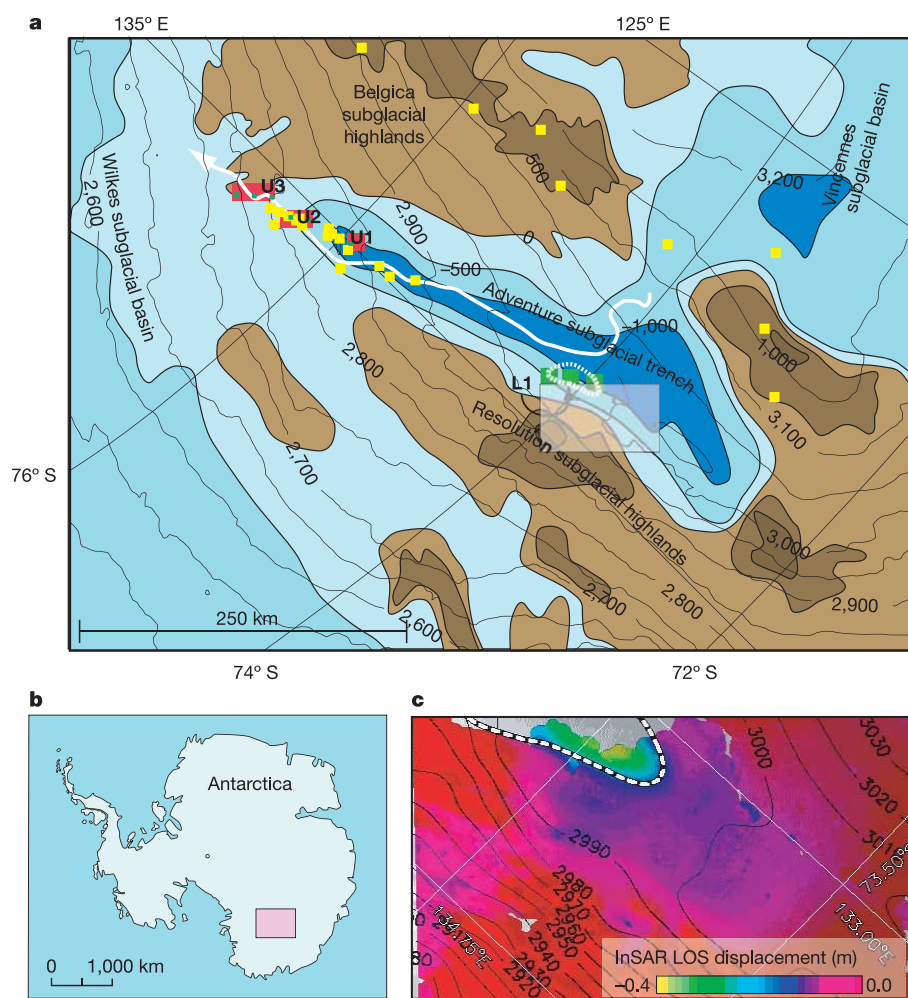


Figure 1 | The origin and flow of an East Antarctic subglacial lake discharge. **a**, Topography of the Adventure subglacial trench region of Dome C, central East Antarctica. The locations of subglacial lakes¹ are denoted by yellow squares and lake L is outlined with a dashed white line. Locations of ERS-2 altimetric data are provided as squares in green ('L' sites; ERS-2 lowering >1 m) and red ('U' sites; ERS-2 uplift >1 m). Ice-surface contours are given as thin lines, and the direction of water flow, calculated from the subglacial water pressure gradient¹⁰, is given by the thick white

arrow. The location of the region is provided in **b**, and the location of the Interferometric Synthetic Aperture Radar (InSAR) data (given in **c**) is provided in the shaded box. Blue and brown shading is included to highlight topography; blue where the bed elevation is below sea level and brown where the bed is higher. **c**, ERS-2 InSAR measurement of ice surface lowering (located in **a**) between 14 August 1997 and 16 March 1998. Ice-surface contours are also shown. LOS, line of site.

unlikely that the pathway consisted of a 'linked-cavity' system¹⁷ because the ice-flow velocity of $<10 \text{ m yr}^{-1}$ precludes the sliding that is required to sustain the cavities of that type of distributed drainage system.

We expect that lakes, fed by conduits or systems of conduits sufficient to carry the basal meltwater, form a normal drainage system under much of east Antarctica. We asked whether rapid discharges such as described here are the normal mode of transfer between the lakes. For a lake to exist, there must be a local reversal of hydraulic gradient. If the lake is filling, this will eventually be overcome, and a hydraulic connection between the lakes becomes possible. When the reservoir is open to the atmosphere, then quasi-periodic rapid discharging is normal⁴ because of the instability of conduit drainage to small changes in its discharge³. However, when the reservoirs are roofed with ice, some of the available energy must be used in deforming the roof. The ratio of the rate of energy release to that of deformation varies as does the ratio of the lakes' areas to their circumference, so discharges between lakes having a small area (or connected with small hydraulic gradient) will be moderated by the motion of the ice roof, possibly to the point of stability.

To get some idea of the magnitudes involved, the deformation energy over lake L may be estimated by assuming the deformation

occurs by a steady, shear strain rate of $\sim 2 \times 10^{-11} \text{ s}^{-1}$ around the $\sim 9 \times 10^4 \text{ m}$ circumference of the lake and that the corresponding stresses may be determined from the flow law of ice¹⁵. Assuming the ice temperature varies linearly from -50°C at the surface, the energy of deformation is $8 \times 10^{13} \text{ J}$ (Supplementary Methods 2). If that of the 'U' sites is similar, the deformation energy is small in comparison with the total available energy, suggesting that the lakes are large enough that their ice roofs had little effect on the dynamics of this flood. (The equating of the total available energy to latent heat of fusion, as was done earlier, also assumes that the energy used in heating water is small. If the water remains near the pressure melting point, this is a good assumption.) But the average depth to lake L is not accurately known. It may in particular be deeper than the BEDMAP-derived¹⁸ value we have used (-285 m above sea level) and in this case the roof deformation may have influenced the shape of the hydrograph in Fig. 2. More theoretical development of the situation is required before a detailed discussion of the hydrograph is possible.

Whatever the details, lake L is evidently unstable, and lakes larger than lake L (or ones connected by larger potential gradients) that are subject to slow filling (the likely case in East Antarctica where basal melt rates¹⁹ are $\sim 1 \text{ mm yr}^{-1}$) will normally⁴ undergo quasi-periodic rapid discharges. While those of lake L will be modest and frequent

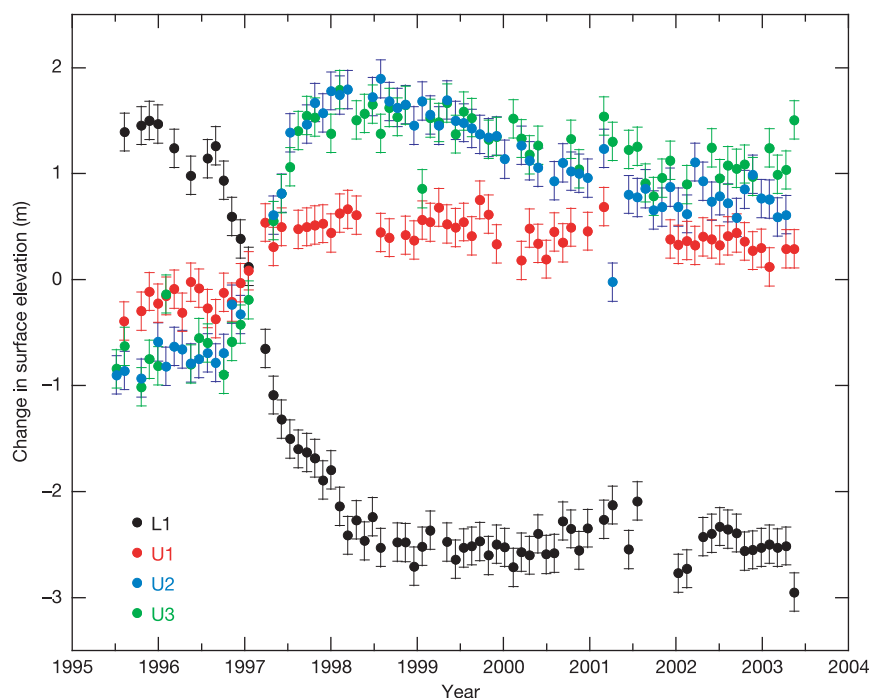


Figure 2 | ERS-2 altimetric data from the four sites, L1, U1, U2 and U3, located in Fig. 1. The error bars are 1-sigma errors, which are determined empirically by examining the error in elevation change of nearby, stationary

ice. Each altimetric data point was averaged over 16 repeat orbit cycles, which leads to an error of ± 0.18 m.

(we estimate from the upstream collecting area a recurrence every 36 years; see Supplementary Methods 3), the volume and periodicity of the discharge depends on the bed geometry, ice cover and basal melt rate, and a spectrum of discharge volumes and intervals are likely. The largest present-day subglacial lake, Lake Vostok, contains $\sim 5,400 \text{ km}^3$ of water²⁰. It is possible that this lake has experienced rapid discharging in the past and, if it is filling, will do so again. The Antarctic Dry Valleys exhibit channels of 10^3 m^2 cross-section that attest to intense, subglacial floods of considerable size⁵ that reached the ice-sheet margins. Similar features are present at the Soya Coast⁶. Together, these coastal records, and our observations of a subglacial lake discharge under the ice-sheet centre, support a common process allowing water stored beneath large ice sheets to escape rapidly to the ocean. Such a process has been hypothesized as responsible for huge discharges from beneath the Laurentide Ice Sheet, which may have been of a magnitude associated with Dansgaard–Oeschger events (A. Fowler, G. Evatt and C. Clark, manuscript in preparation). Whether a large flood may reach the coast from the deep interior will depend on how many downstream lakes lie in its path, and how much energy is required to deform their roofs.

Although the period of such large events is long in comparison with that of lake L, it is short in comparison with the age of the basal ice from which the water derives. Discussions of subglacial lake water residence times^{7,8} typically regard lakes as near-closed systems with slow mass turnovers. It is more likely that the entire drainage system is regularly flushed. Whereas between floods, slow geothermal or pressure-melting forced circulations may establish themselves, the mixing and transport associated with incoming flood water may affect the thermohaline structure and suspended sediment of the lake water column. Periodic discharges of subglacial lake water will reduce the potential for enhanced levels of dissolved gas in the lake, in contrast to that proposed for Lake Vostok under the assumption that it has been hydrologically closed for several million years²¹, and will affect the level of microbial diversity between lakes located in the same flood drainage system. Isotopic analyses of lake water (such as $\delta^{18}\text{O}$ studies²²), that depend on knowing the mass history or source

of the lake water, will also need care. Finally, we note that *in situ* exploration of subglacial Antarctic lakes risks the rapid contamination of significant components of drainage systems.

Received 8 September 2005; accepted 17 February 2006.

1. Siegert, M. J., Carter, S., Tabacco, I., Popov, S. & Blankenship, D. A revised inventory of Antarctic subglacial lakes. *Antarct. Sci.* **17**, 453–460 (2005).
2. Siegert, M. J. *et al.* Physical, chemical and biological processes in Lake Vostok and other Antarctic subglacial lakes. *Nature* **414**, 603–609 (2001).
3. Nye, J. F. Water flow in glaciers, jökulhlaups, tunnels and veins. *J. Glaciol.* **17**, 181–207 (1976).
4. Fowler, A. C. Breaking the seal at Grimsvotn, Iceland. *J. Glaciol.* **45**, 506–516 (1999).
5. Denton, G. E. & Sugden, D. E. Meltwater features that suggest Miocene ice-sheet overriding of the Transantarctic Mountains in Victoria Land, Antarctica. *Geograf. Ann.* **87**, 67–85 (2005).
6. Sawagaki, T. & Hirakawa, K. Erosion of bedrock by subglacial meltwater, Soya Coast, East Antarctica. *Geograf. Ann.* **79**, 223–238 (1997).
7. Kapitsa, A., Ridley, J. K., Robin, G. de Q., Siegert, M. J. & Zotikov, I. Large deep freshwater lake beneath the ice of central East Antarctica. *Nature* **381**, 684–686 (1996).
8. Bell, R. E. *et al.* Origin and fate of Lake Vostok water refrozen to the base of the East Antarctic ice sheet. *Nature* **416**, 307–310 (2002).
9. Siegert, M. J., Taylor, J. & Payne, A. J. Spectral roughness of subglacial topography and implications for former ice-sheet dynamics in East Antarctica. *Glob. Planet. Change* **45**, 249–263 (2005).
10. Shreve, R. L. Movement of water in glaciers. *J. Glaciol.* **11**, 205–214 (1972).
11. Wingham, D. J., Ridout, A., Scharroo, R., Arthern, R. & Shum, C. K. Antarctic elevation change from 1992 to 1996. *Science* **282**, 456–458 (1998).
12. Gray, L. *et al.* Evidence for subglacial water transport in the West Antarctic Ice Sheet through three-dimensional satellite radar interferometry. *Geophys. Res. Lett.* **32**, L03501, doi:10.1029/2004GL021387 (2005).
13. Joughin, I., Kwok, R. & Fahnestock, M. Estimation of ice-sheet motion using satellite radar interferometry: Method and error analysis with application to Humboldt Glacier, Greenland. *J. Glaciol.* **42**, 564–575 (1996).
14. Roberts, M. J. Jökulhlaups: a reassessment of floodwater flow through glaciers. *Rev. Geophys.* **43**, doi:10.1029/2003RG000147 (2005).
15. van der Veen, C. J. *Fundamentals of Glacier Dynamics* 1–462 (Balkema, Rotterdam, 1999).
16. Björnsson, H. Hydrological characteristics of the drainage system beneath a surging glacier. *Nature* **395**, 771–774 (1998).
17. Kamb, B. Glacier surge mechanism based on linked cavity configuration of the basal water conduit system. *J. Geophys. Res.* **92**, 9083–9100 (1987).

18. Lythe, M. B., Vaughan, D. G. & the BEDMAP consortium. *Bedmap–Bed Topography of the Antarctic. 1:10,000,000 map*. Misc. 9 (British Antarctic Survey, Cambridge, 2000).
19. Huybrechts, P. The Antarctic ice sheet and environmental change: a three dimensional modelling study. *Rep. Polar Res.* **99**, 1–241 (Alfred-Wegener-Institut für Polar und Meeresforschung, 1992).
20. Studinger, M., Bell, R. E. & Tikku, A. A. Estimating the depth and shape of subglacial Lake Vostok's water cavity from aerogravity data. *Geophys. Res. Lett.* **31**, L12401, doi:10.1029/2004GL019801(2004).
21. McKay, C. P., Hand, K. P., Doran, P. T., Andersen, D. T. & Priscu, J. C. Clathrate formation and the fate of noble and biologically useful gases in Lake Vostok, Antarctica. *Geophys. Res. Lett.* **30**, doi:10.1029/2003GL017490 (2003).
22. Royston-Bishop, G., Tranter, M., Siegert, M. J., Lee, V. & Bates, P. D. Is Vostok lake in steady state? *Ann. Glaciol.* **39**, 490–494 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature. A figure summarising the main result of this paper is available in Supplementary Information.

Acknowledgements We thank A. Fowler and I. Joughin for their insights and A. Payne and A. Le Brocq for assistance with the calculation of water flow paths. Funding for this work was provided by the NERC Centre for Polar Observation and Modelling.

Author Contributions D.J.W. identified altimetric changes across the study region and undertook the calculations of energy exchange and water flow. M.J.S. placed the altimetric changes in the context of subglacial topography and the locations of known subglacial lakes. A.P.S. processed interferometric synthetic aperture radar data in Fig. 1b. A.S.M. processed ERS-2 altimetric records. D.J.W. and M.J.S. wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.J.W. (djw@cpom.ucl.ac.uk) or M.J.S. (m.j.siegert@bristol.ac.uk).

A Cretaceous terrestrial snake with robust hindlimbs and a sacrum

Sebastián Apesteguía^{1,2} & Hussam Zaher³

It has commonly been thought that snakes underwent progressive loss of their limbs by gradual diminution of their use¹. However, recent developmental and palaeontological discoveries suggest a more complex scenario of limb reduction, still poorly documented in the fossil record^{2–5}. Here we report a fossil snake with a sacrum supporting a pelvic girdle and robust, functional legs outside the ribcage. The new fossil, from the Upper Cretaceous period of Patagonia, fills an important gap in the evolutionary progression towards limblessness because other known fossil snakes with developed hindlimbs, the marine *Haasiophis*, *Pachyrhachis* and *Eupodophis*, lack a sacral region. Phylogenetic analysis shows that the new fossil is the most primitive (basal) snake known and that all other limbed fossil snakes are closer to the more advanced macrostomatan snakes, a group including boas, pythons and colubroids. The new fossil retains several features associated with a subterranean or surface dwelling life that are also present in primitive extant snake lineages, supporting the hypothesis of a terrestrial rather than marine origin of snakes.

*Pachyrhachis problematicus*³, *Haasiophis terrasanctus*⁴ and *Eupodophis descouensi*⁵, three marine fossil snakes from the Tethyan coasts of Northern Gondwana, were until now the only known snakes with well-developed hindlimbs. The presence of fully formed hindlimbs enforced the idea that these were the most primitive (basal) snakes and perfect transitional taxa linking extant snakes to an extinct group of marine lizards, the Mosasauroidae^{3,6,7}. However, the presence of several other features typical of the more advanced macrostomatan snakes such as pythons, boas and colubroids⁸ supports the competing hypothesis that these fossils were advanced (macrostomatan) snakes instead, with no special bearing on the origin of snakes^{4,8,9}. Additionally, all three taxa resemble modern snakes in lacking a sacral region and in having a pelvis that is not suspended from the axial skeleton but rather lies within the ribcage¹⁰.

The snake reported here was found in the context of a rich early Upper Cretaceous fossil fauna^{11,12} from north Patagonia, Argentina, and represents the earliest limbed snake from a fully terrestrial deposit. It retains several primitive features absent in any known fossil or recent snake, including a remarkably primitive pelvis.

Squamata Oppel, 1811

Serpentes Linnaeus, 1758

Najash rionegrina gen. et sp. nov.

Etymology. From Hebrew *Najash*, the legged biblical snake; *rionegrina*, for Río Negro Province, Argentina, where the fossil was found.

Holotype. MPCA (Museo Paleontológico 'Carlos Ameghino', Cipolletti, Río Negro) 390–398, 400, consists of a large fragment of the left dentary and anterior portion of the corresponding splenial, and a nearly complete and articulated postcranial skeleton composed of several sections bearing a total of 122 articulated and

associated vertebrae (109 presacral, 2 sacral, 11 caudals), pelvic girdle, two femora, one fibula and the proximal head of the right tibia (Figs 1 and 2).

Locality and horizon. Upper section of the Candeleros Formation (Cenomanian–Turonian¹³) at 'La Buitrera', Río Negro Province, Argentina.

Additional material. The posterior half of a non-associated braincase with its right otico-occipital region preserved (Fig. 1a) and several associated presacral vertebrae (MPCA 385); several disarticulated cranial and vertebral elements of a larger individual, including an incomplete left dentary, axis, and associated presacral and caudal vertebrae (MPCA 380–383).

Diagnosis. A snake with a strongly concave ventral surface of the parasphenoid rostrum, forming a deep and straight gutter; two sacral vertebrae present; single large parazygantral foramen on each side of neural arch; proximal caudal vertebrae with blunt haemapophyses; robust femora with a large trochanter.

The three specimens referred to *N. rionegrina* are identified as snakes on the basis of the completely enclosed braincase, fused parietals, axis with sutured anterior and fused posterior hypapophyses, large number of presacral vertebrae (109 preserved), zygosphenal and zygapophyseal facets separated by a non-articular area, anterior margin of zygosphenal tectum straight or slightly convex, divided synapophyses, three distally forked lymphapophyses, haemapophyses on posterior tail vertebrae.

The preserved posterior portion of the braincase of *Najash* (Fig. 1a) is similar in several respects to that of *Dinilysia patagonica* and that of the fossorial anilioid snakes (pipesnakes). As in *Dinilysia* and the anilioids, the otico-occipital portion of the braincase is transversely expanded. The right posterodorsal portion of the prootic and posterolateral portion of the parietal form a deep and narrow recess that receives the anterior portion of the missing supratemporal, which was incorporated into the cranial wall as in *Dinilysia* and the anilioids *Cylindrophis* and *Anilius*. The recess is located laterally to the contact between the prootic and the supraoccipital, suggesting a dorsal exposure of the prootic between the supratemporal, exoccipital and supraoccipital, a characteristic also present in *Dinilysia*, *Cylindrophis* and *Anilius*. The laterosphenoid is lacking, a plesiomorphic condition found in *Dinilysia* and scolecophidians (wormsnakes). The contact between the posteriorly expanded edge of the supraoccipital and the right exoccipital suggests that the exoccipitals did not meet dorsally, a plesiomorphic lizard trait also known in *Haasiophis* and in some boine snakes¹⁴. Although present, the crista circumfenestralis of *Najash* (here represented mostly by the crista prootica) is the least developed among all known snakes. The crista prootica projects weakly laterally to the stapedial footplate, around its anterodorsal portion, although without overlapping the latter as in all modern snakes. The ventrolateral expansions of both

¹Museo Argentino de Ciencias Naturales 'Bernardino Rivadavia', A. Gallardo 470, Buenos Aires (1405), Argentina. ²Fundación de Historia Natural 'Félix de Azara' (CEBBAD), Universidad Maimónides, V. Virasoro 732, Buenos Aires (1405), Argentina. ³Museu de Zoologia da Universidade de São Paulo, Av. Nazaré 481, Ipiranga, São Paulo 04263-000, Brazil.

crista interfenestralis and crista tuberalis are broken, but the juxtaposed recess remains widely open posteriorly because of a poorly developed posterodorsal margin of the crista tuberalis, a plesiomorphic feature also found in scolecophidians, *Dinilysia* and basal alethinophidian snakes. The stapedial footplate is broad, as in *Dinilysia* and the fossorial macrostomatid *Xenopeltis*. As in lizards, *Dinilysia* and *Wonambi*, the basiptyergoid processes are prominent, rather longitudinally oriented structures that fit in an articular facet of the pterygoid instead of contacting only the latter as in the more derived snakes. The left dentary retains two mental foramina (Fig. 1b, c). Although no teeth are preserved, their alveoli are transversely expanded instead of anteroposteriorly wide as in macrostomatids. A discrete basal plate is present, but the dentaries lack a lingual ridge (or subdental shelf) medial to the tooth-bearing region, a primitive condition absent from all known snakes.

The cotyle of the procoelous vertebrae form a rounded to slightly oval surface that receives a rounded condyle (Fig. 1d, e), a derived condition shared with *Dinilysia* and alethinophidian snakes, including *Pachyrhachis* and *Haasiophis*. The neural arch is low, as is typical in secretive or fossorial forms, and shares with *Dinilysia* a prominent ridge on each side and above the interzygantral ridge. The zygosphenes are thick and well developed, as in *Dinilysia* and macrostomatids; however, unlike the latter, the interzygapophyseal constriction is shallow and a posterior neural arch notch is absent. Accessory processes of the prezygapophyses are lacking and parazygantral foramina are present in all trunk vertebrae. The anterior trunk

vertebrae bear relatively high and narrow neural spines, developed hypapophyses, and synapophyses directed posteroventrally. Mid-trunk vertebrae are broader, with lower and longer, blade-like neural spines, shallow and thin haemal keels extending along the entire ventral surface of the centrum, and synapophyses reaching the level of the prezygapophyseal tip. Synapophyses are divided into parapophyses and diapophyses, a derived condition shared with all alethinophidian snakes. Conversely, synapophyses project laterally beyond the level of the prezygapophyseal tip on the slightly smaller posterior trunk vertebrae (Fig. 1d, e) and are no longer divided in the more posterior ones. Neural spines become reduced to mere ridges and haemal keels broaden significantly.

Unlike any other snake, *Najash* retains two sacral vertebrae that separate the trunk region anatomically from the caudal region (Fig. 2). Preserved elements of the appendicular skeleton include the pelvic girdle, both femora, the proximal articular head of the right tibia, and the right fibula. Only the right pelvic elements are well preserved and mostly in place. In contrast with *Haasiophis*, *Pachyrhachis* and *Eupodophis*, the last rib in *Najash* is ventral to the right femur, reflecting the external position of the hindlimbs with respect to the ribcage (Fig. 2b). Sacral pleurapophyses are long and slightly curved, and their pointed tips are separated, suggesting a loose suspension of the pelvis. Pubis, ilium and ischium are not sutured or fused together proximally, and the medioventral pubo-ischiac symphysis is lacking. Both ilium and pubis are similar in size and show a rounded, expanded proximal head and a long rod-like

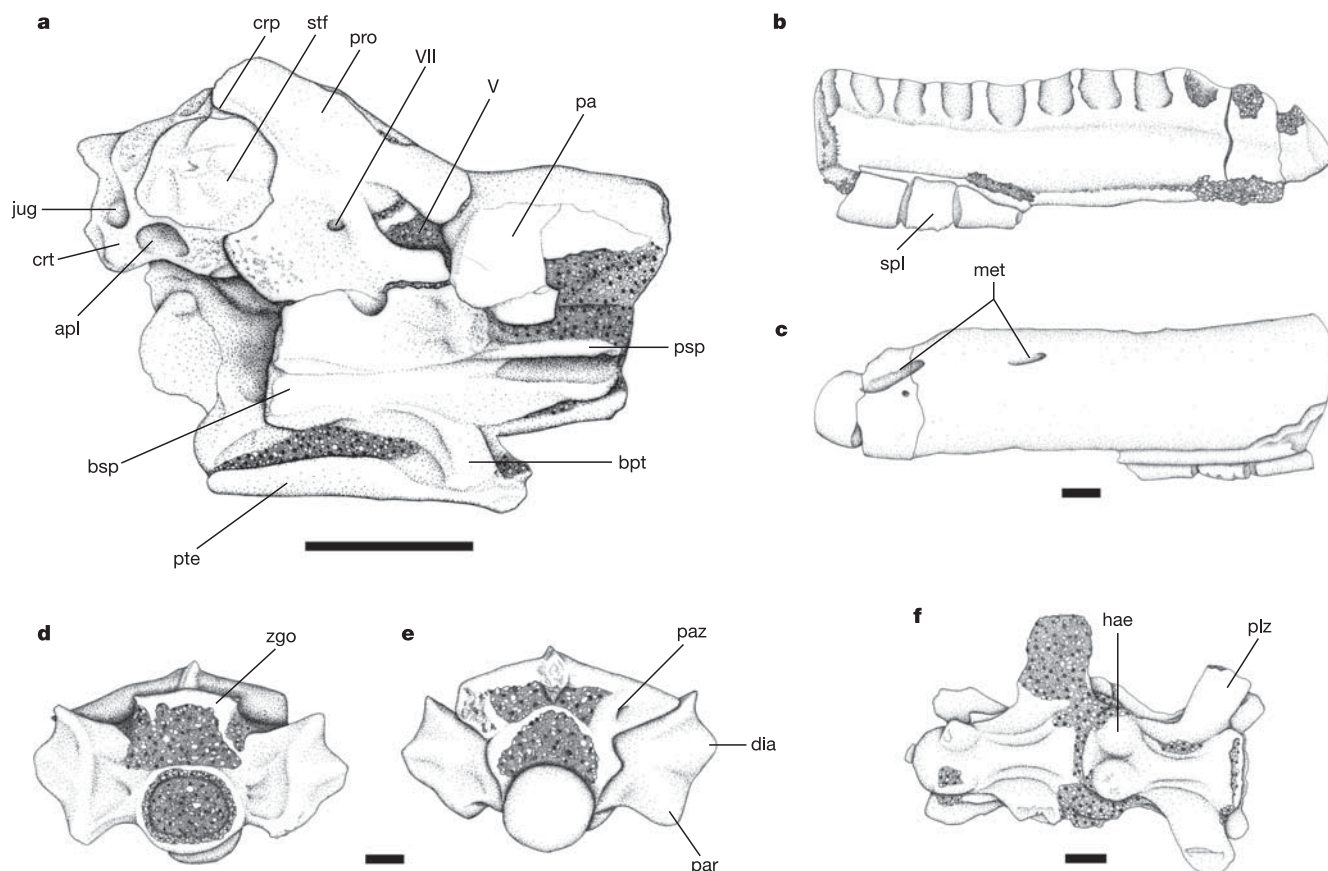


Figure 1 | *Najash rionegrina*. **a**, Posterior half of the braincase in right ventro-lateral view (MPCA 385). **b**, **c**, Left dentary and splenial of the holotype (MPCA 390–398) in medial (**b**) and lateral (**c**) views. **d**, **e**, Posterior trunk vertebra of the holotype (MPCA 390–398) in anterior (**d**) and posterior (**e**) views. **f**, Two articulated caudal vertebrae of the holotype (MPCA 390–398) in ventral view. Scale bars, 5 mm (**a**), 1 mm (**b–f**). Abbreviations: apl, lateral opening of recessus scalae tympani; bpt,

basiptyergoid process; bsp, basisphenoid; crp, crista prootica; crt, crista tuberalis; dia, diapophysis; hae, haemapophysis; jug, jugular foramen; met, mental foramen; par, parapophysis; pa, parietal; paz, parazygantral foramen; pro, prootic; plz, caudal pleurapophysis; psp, parasphenoid; pte, pterygoid; spl, splenial; stf, stapedial footplate; zgo, zygosphenes; V, foramen for the maxillary and mandibular branches of the trigeminal nerve; VII, foramen for hyomandibular branch of facial nerve.

body that tapers distally. The right ischium is broken in two pieces. However, the proximal part retained its contact with the two other pelvic elements. As in *Pachyrhachis*, it is half the size of the latter two pelvic bones and is spatula-shaped proximally. The first three caudal vertebrae bear well-defined, distally bifurcated lymphapophyses that project laterally. Free lymphapophyses (bifurcated ribs) and chevron bones are lacking. Haemapophyses are present as small button-like nodules on the posterior edge of the centrum of the more posterior caudal vertebrae (Fig. 1f), resembling the condition found in anilioids and some macrostomatan snakes.

A phylogenetic analysis, including all relevant fossil snakes, shows *Najash* as the most basal snake (Fig. 3), lying outside the clade consisting of all living snakes. Statistical support for this hypothesis is strong. All other mid-Cretaceous snakes are nested within the clade formed by living snakes, with the terrestrial *Dinilysia* as the sister-group of alethinophidians, whereas the marine *Haasiophis*, *Pachyrhachis* and *Eupodophis* as well as the Pleistocene *Wonambi naracoortensis* are nested within a poorly resolved macrostomatan clade, supporting a macrostomatan affinity^{4,8–10,15}, instead of a basal position as the most primitive snakes^{3,6,7}.

Both cranial (a transversely expanded occipital region and a broad stapedial footplate) and vertebral (a low neural arch) morphological traits of *Najash* show adaptations to a subterranean life, perhaps as a surface-dwelling species that would occasionally use tunnels

produced by burrowers. *Najash*, scolecophidians, *Dinilysia* and anilioids represent the four successive outgroups to the macrostomatan clade in which the first marine snakes were documented. This scenario unequivocally supports the hypothesis of a subterranean or surface-dwelling origin of snakes.

METHODS

The data matrix used in the phylogenetic analysis is based on two recently published character lists^{4,15}. Twenty-one new characters were added to include post-cranial morphology, totalling 119 characters coded for 18 snake taxa. Character codings for the fossil snakes *Wonambi naracoortensis*, *Pachyrhachis problematicus* and *Eupodophis descouensi*, considered by some authors as being the most basal snakes⁷, were reviewed in accordance with recently published descriptions^{16–18}. Codings for *Dinilysia patagonica* are based on observations made by H.Z. on the holotype and new specimens recently discovered¹⁹. Characters included in this matrix are intended to address the more inclusive basal snake interrelationships and the affinities of the relevant fossils *Najash*, *Pachyrhachis*, *Haasiophis*, *Eupodophis* and *Wonambi*; and does not attempt to elucidate macrostomatan interrelationships. Recent attempts to reconstruct phylogenetic affinities within the macrostomatan clade show extensive conflicting character delimitations^{3,4,7–10,14,15} that should be addressed more thoroughly before any new proposal of macrostomatan interrelationships. We rooted the analysis with a hypothetical varanoid ancestor (coded according to the conditions found in the terrestrial *Varanus*, *Heloderma* and *Lanthanotus*, and the marine Mosasauroidea⁷). We alternatively used a dibamid ancestor as an out-group⁹ with no effects relevant to this study. Analyses were performed with

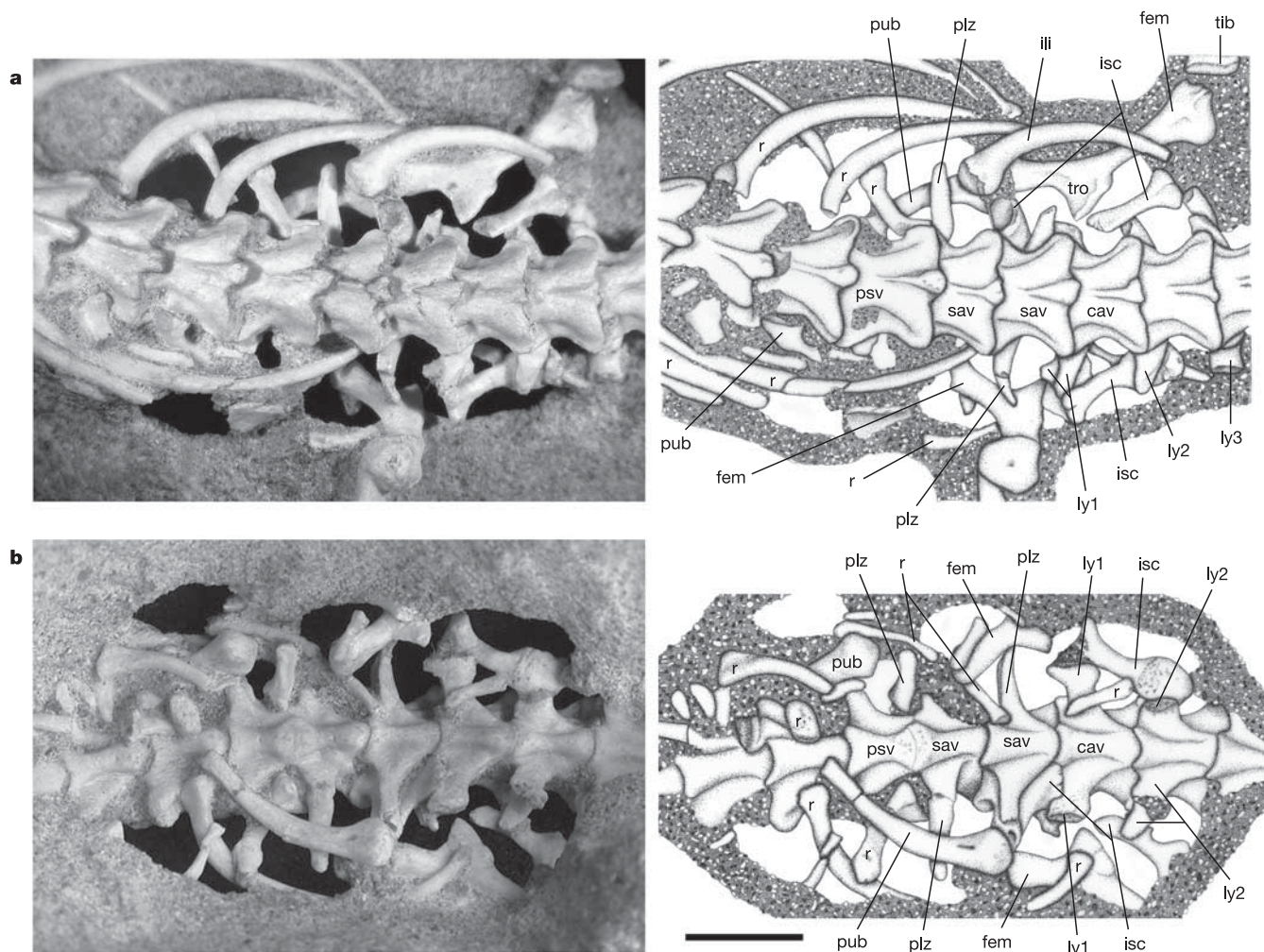


Figure 2 | Sacral region of the holotype of *Najash rionegrina* (MPCA 390–398). a, Dorsal view. b, Ventral view. The left pelvic and limb elements show signs of healed traumatism, with a large callus formation on the fractured femur. The disarticulated fibula is not visible in these views. Scale

bar, 50 mm. Abbreviations: cav, first caudal vertebra; fem, femur; ili, ilium; isc, ischium; ly1–ly3, first, second, and third lymphapophyses; plz, sacral pleurapophysis; pub, pubis; r, rib; sav, sacral vertebrae; tib, tibia; tro, trochanter; psv, last presacral vertebra.

Evolution of cooperative strategies from first principles

Mikhail Burtsev¹ & Peter Turchin²

One of the greatest challenges in the modern biological and social sciences is to understand the evolution of cooperative behaviour. General outlines of the answer to this puzzle are currently emerging as a result of developments in the theories of kin selection^{1–7}, reciprocity^{8–10}, multilevel selection^{11–15} and cultural group selection^{16,17}. The main conceptual tool used in probing the logical coherence of proposed explanations has been game theory, including both analytical models and agent-based simulations^{6,7,9,18–24}. The game-theoretic approach yields clear-cut results but assumes, as a rule, a simple structure of payoffs and a small set of possible strategies. Here we propose a more stringent test of the theory by developing a computer model with a considerably extended spectrum of possible strategies. In our model, agents are endowed with a limited set of receptors, a set of elementary actions and a neural net in between. Behavioural strategies are not predetermined; instead, the process of evolution constructs and reconstructs them from elementary actions. Two new strategies of cooperative attack and defence emerge in simulations, as well as the well-known dove, hawk and bourgeois strategies. Our results indicate that cooperative strategies can evolve even under such minimalist assumptions, provided that agents are capable of perceiving heritable external markers of other agents.

The two-dimensional artificial world in our model is divided into cells, which either contain a resource bundle or are empty. An empty cell can acquire a resource bundle with a certain probability per time step and lose it when resource is consumed by an agent. Agents are characterized by a set of receptors, a set of effectors and a neural net connecting receptors to effectors. Each effector is responsible for a particular action. Agents can do nothing (rest), consume the resource bundle if it is present (eat), produce offspring (divide), go forward to a neighbour cell (move), make a turn to the left or right (turn), and attack another agent if present in the same cell (attack). All actions spend energy taken from the agent's internal store. If internal energy is completely depleted, the agent dies. The least energetically demanding action is rest, the most demanding is attack. Consumption of resource increases the internal store of energy subject to an upper limit (the maximum energy that can be stored). When an agent divides, one offspring is created and placed in the same cell as the parent. The parent then transfers half of its energy to the offspring. When one agent hits another, the victim loses an amount of energy, which is gained by the attacker (see Methods for energetic costs of actions).

Sensory inputs of agents include its internal store of energy, whether there are resources in the agent's field of vision (the cell it is in, the neighbour cell in front of the agent, and the cells on the right and left), and how many other agents are in the field of vision. Each agent has external phenotype that is coded by a vector of integer values (markers). The markers do not influence behaviour but function only as indicators of similarity. The euclidian distance

between an agent's markers and the markers of another agent in the cell (a potential subject for attack) is also a sensory input. Behaviour of an agent is controlled by a simple one-layer neural net. Both weights of the neural net and external markers are inherited by the offspring when an agent divides, subject to a set rate of mutation. Details of the implementation of the model are given in the Methods.

All of our simulations were started with an initial population of agents that were unaware of markers (the matrix coefficients connecting input from markers to actions were preset to zero). Thus, the use of markers in a population had to evolve from a blank slate. Because markers and behaviours are not linked (apart from both being inherited from the ancestors), agents can lose cooperative behaviours by mutation while retaining 'in-group' markers. Thus, the structure of the model allows free-riders to arise.

The number of potential behavioural strategies in our model is astronomical ($>10^{1,000}$; see Methods). Because of the vast number of potential strategies, it is difficult to understand exactly what are the behaviours that evolve in the simulation—each matrix of neural weights is a 'black box'. To make sense of our results, we confronted agents that evolved in our simulations with a discrete set of stimuli and noted the action taken. This approach enabled us to classify strategies into aggressive or not, and those discriminating in-group versus out-group members (see Methods).

Our study examined the spectrum of strategies evolving in the full model versus the simplified version in which agents could not detect external markers. We also determined how the carrying capacity of the environment (varied by increasing the size of resource bundles while keeping the rate of bundle appearance constant) affects the strategies evolved.

Analysis of the model without markers showed that the strategies evolving in the simulation corresponded to those in the well-known game of dove–hawk–bourgeois²⁵. Doves never attack other agents and attempt to escape when attacked, whereas hawks make a living by predation on other agents. The bourgeois strategy in our model is to stay in the same cell and immediately attack any invader, while ignoring agents in neighbouring cells (unlike the hawks). In the model without markers the dominant strategy is bourgeois, provided that the carrying capacity of a single cell is sufficient for supporting a sedentary agent (amount of resource in a bundle is sufficient for survival until the next bundle appears). Below this threshold, C_1 , the bourgeois strategy is impossible, because agents are forced to keep moving to get enough food to survive, and the population is divided between doves and hawks. The long-term population density increases linearly with carrying capacity until it reaches C_1 and then becomes flat (Fig. 1). This is because once the bourgeois strategy takes over, each cell can be occupied by only a single agent. Even when resources are sufficient to support more than one agent per cell, they fight until only one remains.

¹Department of Non-linear Dynamics, Keldysh Institute of Applied Mathematics of RAS, Moscow 125047, Russia. ²Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut 06269-3042, USA.

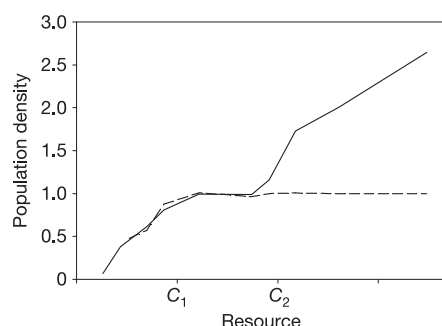


Figure 1 | The effect of resource abundance on population density.

Population density (average number of agents per cell) as a function of the abundance of resources (the size of resource bundle randomly appearing in a cell with a fixed probability). Broken curve, model version without external markers; solid curve, model with markers.

In the full model, in which agents can evolve the ability to detect phenotypic similarity, three kinds of cooperative strategies emerged. The first one was simply the cooperative version of the dove. Cooperative doves ignored out-group (phenotypic distance large) members, but left cells with in-group (phenotypic distance is small) members to avoid competing with them. In the second strategy, agents also left cells with in-group members, but when they detected out-group members they attacked them. We term this strategy 'raven', because, according to the Russian proverb, 'a raven will not peck out the eye of another raven'. The third cooperative strategy was to stay in the same cell with in-group members and collectively fight with any out-group invader. Having to share limited resources of the cell meant that agents using this cooperative defence strategy were small (had small stores of internal resource), but they still had a good chance of defeating a large invader because of their advantage in numbers. This strategy resembles the 'mobbing' behaviour that many species of small birds, such as starlings, use to drive away large predators. For this reason, we call it the 'starling' strategy.

The emergence of the starling strategy has a marked effect on the relationship between carrying capacity and long-term population density. For lower values of carrying capacity, the curve in the full model follows that of the model without markers. But once it exceeds the threshold $C_2 = 2C_1$, the density curve again begins to rise (Fig. 1). Analysis of the effect of carrying capacity on the prevalent strategies, evolved by agents, indicates the mechanism of this rise. If carrying capacity is insufficient to support at least two agents in a cell ($C < C_2$), the starling strategy cannot invade the population. Instead, the dominant strategy is raven, whose frequency increases linearly with C for $C < C_1$, and saturates at a high level for $C_1 < C < C_2$ (Fig. 2). So far, the only difference resulting from

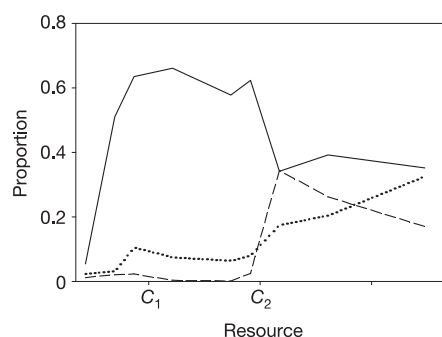


Figure 2 | The effect of resource abundance on the long-term proportions of agents using various strategies. Average proportion of agents using the raven (unbroken line), the cooperative dove (dashed line) and the starling (dotted line) strategies in the full model with markers as a function of the abundance of resources. Proportions do not add up to one because there are other strategies, including the non-cooperative ones.

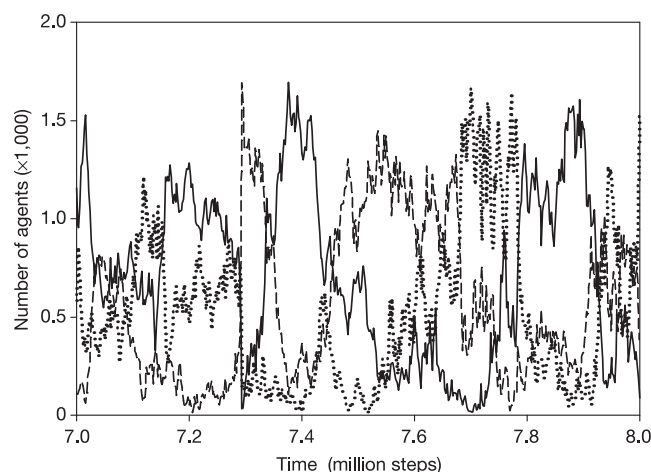


Figure 3 | Dynamic coexistence of the raven, the cooperative dove and the starling strategies. Numbers of agents using the raven (dashed line), the cooperative dove (dotted line) and the starling (unbroken line) strategies as a function of time in one realization of the model.

the agents' ability to use markers is the switch from the hawk and bourgeois to the raven strategy (predators recognize and do not attack in-group members). Once carrying capacity goes beyond C_2 , the starling strategy becomes viable. Starlings, however, do not drive ravens to extinction. Instead, we observe sustained oscillations in the numbers of starlings, ravens and doves (Fig. 3). As a result, all three strategies manage to coexist in the long term (Fig. 2).

For low carrying capacity ($C < C_1$), we observed the emergence of another new strategy, which was to leave the cell whenever any other agent (related or not) appeared in it. This strategy differs from that of the doves, who ignore unrelated agents and escape only if attacked. In fact, it is the exact opposite of the bourgeois and resembles strategies predicted to evolve at the 'anti-private-property equilibrium' (H. Gintis, personal communication). Another interesting behaviour was observed in starlings, which apparently show the previously predicted 'desperado effect'²⁶. If an internal resource of one of starlings in a cell falls below a certain level, this agent leaves the cell—apparently preferring to face an almost certain death in a fight for resources with out-group members, rather than continuing to deplete a possibly inadequate resource in its native cell.

Our results have important implications for the evolution of territoriality in animals (and private property in humans). With a few exceptions²⁷, theorists have paid little attention to the role that cooperation may have in the evolution of territoriality²⁸. Our study suggests that cooperative defence of territory can radically change the

Table 1 | List of input variables and their definitions

Input variable*	Value
I_1	Bias constant, k
I_2, I_3, I_4, I_5	k if there is resource bundle in the field of agent's vision; 0 in the opposite case
I_6, I_7, I_8, I_9	cN_c , where c is a constant, N_c is the number of agents in the given cell of the field of agent's vision
I_{10}	Value of internal resource, r
I_{11}	$r_{\max} - r$
I_{12}	$\sqrt{\sum_i (\bar{m}_i - \bar{m})^2}$, where $\bar{m}$ is a centroid of markers of all agents at the current cell
I_{13}	$k \cdot \sqrt{\sum_i (m_i^p - m_i)^2} / 2M_{\max}$, where m^p is a marker of partner to interact

* Note that I_1 is a constant and that I_2 to I_5 are binary variables (k is a functional analogue of unity and was set equal to $r_{\max}$, where $r_{\max}$ is the maximal possible value of stored internal resource).

Table 2 | The energetic costs of an agent's actions*

Output vector	Action	Change of internal resource $r_i^\dagger$
O_0	Rest	$-0.001r_{\max}$
O_1	Turn left	$-0.002r_{\max}$
O_2	Turn right	$-0.002r_{\max}$
O_3	Consume the resource bundle	$+[(0.06r_{\max}, 0.4r_{\max})^\ddagger]$
O_4	Move	$-0.004r_{\max}$
O_5	Divide	$-0.004r_{\max}\S$
O_6	Fight (randomly chosen agent in the cell)	The cost of attack is $0.1r_{\max}$; the gain is $+0.2r_{\max}$ if internal resource of the victim is $r_n \geq 0.2r_{\max}$ and $+r_n$ otherwise; the victim loses $-0.2r_{\max}$

*Note that $r_{\max}$ is the energy storage capacity.

† This scheme of setting parameter values reflects our assumption that the energetic cost of movement (move, turn left, and so on) is greater than the cost of resting, whereas the cost of attack is much greater than the cost of movement. Note that energetic losses are indicated with a minus sign and gains with a plus sign.

‡ The energy intake was a parameter in the series of simulations in which food appears in the cell with the constant probability of 0.01 and the amount of resource in the bundle was varied between $0.06r_{\max}$ and $0.4r_{\max}$.

$\S$ When the agent divides it spends $0.004r_{\max}$; half of the remaining energy is then transferred to the offspring.

course of evolution in resource-rich ($C > C_2$) environments. When the amount of resource becomes large enough to support more than one agent, and too large for a single agent to monopolize, solitary bourgeois are replaced by cooperative starlings, provided that agents can recognize in-group members. The starling strategy does not take over completely, however, but coexists with other strategies in a complex dynamical way (Fig. 3).

One potential strategy that did not evolve in our simulations was cooperative attack (the 'wolf' strategy), probably because agents lacked the appropriate effectors for travelling in groups in search of prey. In future work we plan to investigate whether adding such actions as 'follow another agent' could allow evolution of cooperative predation. Another limitation of our study was that agents could transmit traits (including phenotypic ones) only vertically from parent to offspring. This means that our 'in-group members' were also relatives. But one of the greatest puzzles about human ultrasociality is how cooperation between unrelated individuals can arise in the process of evolution. This issue can be addressed (and we plan to do so) by allowing cultural transmission of traits between group members.

In conclusion, our study shows that within the artificial evolution framework it is possible to model not only how one strategy displaces another (or not), but the very process by which new strategies emerge out of a very large space of possibilities. Our model did not endow agents with a set of preconceived strategies—all that we assumed was that agents have a set of elementary sensory inputs and a set of actions. The selection of appropriate connections between inputs and actions was moulded by the process of evolution. It is notable that the agents in our simulations evolved many of the strategies that were postulated by previous researchers. Thus, in the absence of phenotypic markers, three distinct strategies emerged corresponding to the dove, the hawk and the bourgeois. This shows that our results are not in opposition to game theory, but represent an extension of previous approaches. In the presence of markers, the evolution resulted in some predictable modifications of these basic strategies, but also in the emergence of a new one. Cooperative doves avoided competition with in-group members, whereas cooperative hawks—'ravens'—avoided attack on phenotypically similar agents. The new strategy was the starlings, who lived in groups and defended territory cooperatively against predation.

METHODS

Agents' behaviour and evolution. Behaviour of agents is governed by a simple control system in which each output associated with a specific action is connected to sensory inputs from the environment or the internal state of the

agent. The control system is linear and functions similarly to a feed-forward neural network with no hidden layer. To calculate the output vector $\mathbf{O}$ of values, the input vector $\mathbf{I}$ is multiplied by a matrix of weights $\mathbf{W}$, which are constrained to lie in the range $[-W_{\max}, W_{\max}]$:

$$O_j = \sum_i w_{ij} I_i \quad (1)$$

At each time step, the agent performs the action associated with the maximum output value (note that the order in which agents act is randomly shuffled every step). The input vector $\mathbf{I}$ is populated with information about the presence of resource and other agents in the field of vision (the cell where the agent is, the neighbour cell in front of the agent, and the cells on the right and left), the level of internal resource and the euclidean distance between marker vectors of the agent and its partner for potential interaction. A full list of input variables and their definitions are given in Table 1. At the start of simulation, an initial population was formed from the agents with the same matrix of weights $\mathbf{W}$. All the weights in this matrix were set to zero except for three that defined the following simple strategy: move if a resource bundle is in the forward cell; eat if a resource is in the current cell; divide otherwise. Correspondence between outputs and actions, and how changes of the internal resource r depend on actions, are summarized in Table 2.

To speed up simulations, all variables were integers. For all simulations, the size of the world was 900 cells, $W_{\max}$ was 1,000, $r_{\max}$ was 5,000, the dimension of the marker vector was 10, and its values were bounded by $[-W_{\max}, W_{\max}]$.

If the agent executes the action 'divide', its offspring is placed in the same cell. The genome of the offspring is constructed in the following way: first, for every weight of the control system, a random value uniformly distributed on the interval $[-0.03W_{\max}, 0.03W_{\max}]$ is added; second, for every component of the marker, a random value uniformly distributed on the interval $[-0.15W_{\max}, 0.15W_{\max}]$ is added. A preliminary version of the model has been investigated in ref. 29.

Definition of strategies. In our model, every agent has 11 independent inputs from 13 available (I_1 is constant and $I_{11} = r_{\max} - I_{10}$) and seven actions. Assuming that we take into consideration only two possible values per input (this procedure gives us a lower bound on the estimate), the overall number of strategies can be estimated as $2^{7 \times 11}$, which approximately equals $10^{1.730}$. To reduce such a large space of strategies, we consider only six situations in which the agent was allowed to interact with an in-group or out-group member for three levels of internal resource ($0.02r_{\max}$, $0.5r_{\max}$ and $0.98r_{\max}$). We also group 'rest', 'eat' and 'turn' actions together because they correspond to the absence of direct interaction between agents. As a result, the strategy space was reduced to $4^6 = 4,096$ possible strategies. Frequencies of strategies in the population at a given point in time were calculated by picking every agent and calculating its actions for every situation.

Our classification of strategies, which evolved in the simulation, was based on this subset of strategy space. We treated an agent as a 'raven' if for any value of its internal resource it fights an out-group agent but leaves the cell with an in-group agent. A 'starling' was defined as an agent that does not leave the cell in the presence of in-group individuals and fights out-group agents for any value of internal resource $r \leq 0.5r_{\max}$.

Received 18 July; accepted 22 November 2005.

- Hamilton, W. D. The genetical evolution of social behaviour. I and II. *J. Theor. Biol.* **7**, 1–52 (1964).
- Maynard Smith, J. Kin selection and group selection. *Nature* **201**, 1145–1147 (1964).
- Michod, R. E. & Abugov, R. Adaptive topography in family-structured models of kin selection. *Science* **210**, 667–669 (1980).
- Lacy, R. C. & Sherman, P. W. Kin recognition by phenotype matching. *Am. Nat.* **121**, 489–512 (1983).
- Queller, D. C. A general model for kin selection. *Evolution* **46**, 376–380 (1992).
- Riolo, R. L., Cohen, M. D. & Axelrod, R. Evolution of cooperation without reciprocity. *Nature* **414**, 441–443 (2001).
- Axelrod, R., Hammond, R. A. & Grafen, A. Altruism via kin-selection strategies that rely on arbitrary tags with which they coevolve. *Evolution* **58**, 1833–1838 (2004).
- Trivers, R. The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57 (1971).
- Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577 (1998).
- Bowles, S. & Gintis, H. The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theor. Popul. Biol.* **65**, 17–28 (2004).
- Sober, E. & Wilson, D. S. *Unto Others: The Evolution and Psychology of Unselfish Behavior* (Harvard Univ. Press, Cambridge, Massachusetts, 1998).
- Wilson, D. S. Altruism and organism: disentangling the themes of multilevel selection theory. *Am. Nat.* **150** (suppl.), 122–134 (1997).

13. Gintis, H. Strong reciprocity and human sociality. *J. Theor. Biol.* **206**, 169–179 (2000).
14. Bowles, S. in *Social Dynamics* (eds Durlauf, S. N. & Young, H. P.) 155–190 (MIT Press, Cambridge, Massachusetts, 2001).
15. Turchin, P. *Historical Dynamics: Why States Rise and Fall* (Princeton Univ. Press, Princeton, New Jersey, 2003).
16. Boyd, R. & Richerson, P. J. *Culture and the Evolutionary Process* (Univ. Chicago Press, Chicago, Illinois, 1985).
17. Richerson, P. J. & Boyd, R. in *Indoctrinability, Ideology, and Warfare* (eds Eibl-Eibesfeldt, I. & Salter, F. K.) 71–95 (Berghahn, New York, 1998).
18. Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
19. Axelrod, R. *The Complexity of Cooperation: Agent-based Models of Competition and Collaboration* (Princeton Univ. Press, Princeton, New Jersey, 1997).
20. Epstein, J. M. & Axtell, R. *Growing Artificial Societies: Social Science from the Bottom Up* (Brookings Institution Press, Washington DC, 1996).
21. Nowak, M. A. & May, R. M. Evolutionary games and spatial chaos. *Nature* **359**, 826–829 (1992).
22. McNamara, J. M., Barta, Z. & Houston, A. I. Variation in behavior promotes cooperation in the Prisoner's Dilemma game. *Nature* **428**, 745–748 (2004).
23. Nowak, M., Sasaki, A., Taylor, C. & Fudenberg, D. Emergence of cooperation and evolutionary stability in finite populations. *Nature* **428**, 646–650 (2004).
24. Hauert, C. & Doebeli, M. Spatial structure often inhibits the evolution of cooperation in the snowdrift game. *Nature* **428**, 643–646 (2004).
25. Maynard Smith, J. The theory of games and the evolution of animal conflicts. *J. Theor. Biol.* **47**, 209–222 (1974).
26. Grafen, A. The logic of divisively asymmetric contests: respect for ownership and the desperado effect. *Anim. Behav.* **35**, 462–467 (1987).
27. Matthiopoulos, J., Moss, R. & Lambin, X. The kin facilitation hypothesis for red grouse population cycles: territorial dynamics of the family cluster. *Ecol. Model.* **147**, 291–307 (2002).
28. Maher, C. R. & Lott, D. F. A review of ecological determinants of territoriality within vertebrate species. *Am. Midl. Nat.* **143**, 1–29 (2000).
29. Burtsev, M. S. Tracking the trajectories of evolution. *Artif. Life* **10**, 397–411 (2004).

Acknowledgements We thank P. Taylor, H. Gintis and C. Cioffi-Revilla for comments on the manuscript. M.B. was supported by the Russian Fund for Basic Research. P.T. was supported by the NSF.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.B. (mbur@narod.ru).

DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage

Michael C. Zody¹, Manuel Garber¹, David J. Adams², Ted Sharpe¹, Jennifer Harrow², James R. Lupski³, Christine Nicholson², Steven M. Searle², Laurens Wilming², Sarah K. Young¹, Amr Abouelleil¹, Nicole R. Allen¹, Weimin Bi³, Toby Bloom¹, Mark L. Borowsky¹, Boris E. Bugalter¹, Jonathan Butler¹, Jean L. Chang¹, Chao-Kung Chen², April Cook¹, Benjamin Corum¹, Christina A. Cuomo¹, Pieter J. de Jong⁴, David DeCaprio¹, Ken Dewar^{1†}, Michael FitzGerald¹, James Gilbert², Richard Gibson², Sante Gnerre¹, Steven Goldstein⁵, Darren V. Grafham², Russell Grocock², Nabil Hafez¹, Daniel S. Hagopian¹, Elizabeth Hart², Catherine Hosage Norman¹, Sean Humphray², David B. Jaffe¹, Matt Jones², Michael Kamal¹, Varsha K. Khodiyar⁶, Kurt LaButti¹, Gavin Laird², Jessica Lehoczy¹, Xiaohong Liu¹, Tashi Lokyitsang¹, Jane Loveland², Annie Lui¹, Pendexter Macdonald¹, John E. Major^{1†}, Lucy Matthews², Evan Mauceli¹, Steven A. McCarroll¹, Atanas H. Mihalev¹, Jonathan Mudge², Cindy Nguyen¹, Robert Nicol¹, Sinéad B. O'Leary¹, Kazutoyo Osoegawa⁴, David C. Schwartz⁵, Charles Shaw-Smith², Pawel Stankiewicz³, Charles Steward², David Swarbreck², Vijay Venkataraman¹, Charles A. Whittaker^{1†}, Xiaoping Yang¹, Andrew R. Zimmer¹, Allan Bradley², Tim Hubbard², Bruce W. Birren¹, Jane Rogers², Eric S. Lander¹ & Chad Nusbaum¹

Chromosome 17 is unusual among the human chromosomes in many respects. It is the largest human autosome with orthology to only a single mouse chromosome¹, mapping entirely to the distal half of mouse chromosome 11. Chromosome 17 is rich in protein-coding genes, having the second highest gene density in the genome^{2,3}. It is also enriched in segmental duplications, ranking third in density among the autosomes⁴. Here we report a finished sequence for human chromosome 17, as well as a structural comparison with the finished sequence for mouse chromosome 11, the first finished mouse chromosome. Comparison of the orthologous regions reveals striking differences. In contrast to the typical pattern seen in mammalian evolution^{5,6}, the human sequence has undergone extensive intrachromosomal rearrangement, whereas the mouse sequence has been remarkably stable. Moreover, although the human sequence has a high density of segmental duplication, the mouse sequence has a very low density. Notably, these segmental duplications correspond closely to the sites of structural rearrangement, demonstrating a link between duplication and rearrangement. Examination of the main classes of duplicated segments provides insight into the dynamics underlying expansion of chromosome-specific, low-copy repeats in the human genome.

Human chromosome 17 is implicated in a wide range of human genetic diseases. It is home to genes involved in early-onset breast cancer (*BRCA1*), neurofibromatosis (*NF1*) and the DNA damage response (*TP53* encoding the p53 protein). The complex rearrangement and duplication structure of chromosome 17 predisposes it to non-allelic homologous recombination (NAHR), resulting in DNA rearrangements that cause several well-studied microdeletion disorders. These include hereditary neuropathy with pressure palsies

(HNPP)⁷ at 17p12, and Smith–Magenis syndrome (SMS) deletions at 17p11.2 (refs 8, 9). Microduplication counterparts for both of these conditions are also known, famously in the case of Charcot–Marie–Tooth disease type 1A (CMT1A) at 17p12 (refs 10, 11), the most common inherited peripheral neuropathy (see ref. 12 for a review), and more recently for SMS¹³. Further, the complex architecture in this region, consisting of ~50-kb subunits of direct and inverted repeats able to form cruciform structures, is also responsible for susceptibility to one of the most common somatic rearrangement events characterized, isodicentric 17q, which is associated with several cancers and signifies poor prognosis¹⁴. Finally, all cases of Miller–Dieker syndrome¹² and many cases of isolated lissencephaly sequence¹² are associated with haploinsufficiency effects of deletions at 17p13.3.

As part of the Human Genome Project², we generated a finished sequence of human chromosome 17, which comprises ~2.8% of the euchromatic genome. The finished sequence contains 78,839,971 bases and is interrupted by nine euchromatic gaps and one gap containing the centromere region (Fig. 1). The total size of the euchromatic gaps is estimated at 854 kb (see Methods and Supplementary Table S1). The frequency and size of gaps is slightly higher than the human genome average². (It is also significantly higher than for mouse chromosome 11; see below.) This is largely due to segmentally duplicated or unclonable regions on human chromosome 17 (see Supplementary Information). A clone list for the most recent version of the sequence is provided in Supplementary Table S2.

An overview of chromosome 17 is shown in Fig. 1. The chromosome has features characteristic of gene-rich regions², including a high average G+C content (45.5%) and a collection of transposable

¹Broad Institute of MIT and Harvard, 7 Cambridge Center, Massachusetts 02142, USA. ²The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. ³Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. ⁴BACPAC Resources, Children's Hospital Oakland Research Institute, 747 52nd Street, Oakland, California 94609, USA. ⁵Laboratory for Molecular and Computational Genomics, University of Wisconsin-Madison, 425 Henry Mall, Madison, Wisconsin 53706, USA. ⁶HUGO Gene Nomenclature Committee, The Galton Laboratory, Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK. [†]Present addresses: McGill University and Genome Quebec Innovation Centre, Montreal, Quebec H3A 1A4, Canada (K.D.); Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, USA (J.E.M.); MIT Center for Cancer Research, 77 Massachusetts Avenue E18-570, Cambridge, Massachusetts 02139, USA (C.A.W.).

element fossils (45.3% overall) with a relative excess of short interspersed elements (SINEs, 22.3%) and a deficit of long interspersed elements (LINEs, 14.4%). It also has a large excess of segmental duplications (defined as having >90% identity and being >1 kb in length¹⁵). Such segmental duplications constitute 8.6% (7 Mb) of the finished sequence, in contrast to a median of 3.9% for human chromosomes and an average of 4.3% across autosomal sequences. As with other duplication-rich chromosomes, chromosome 17 has a large proportion of intrachromosomal duplication: 62.3% of the duplicated sequence is strictly intrachromosomal, 17.4% is both intra- and inter-chromosomal, and 20.3% is solely interchromosomal. We elaborate below on the segmental duplication content of the chromosome.

We have produced a manually curated catalogue of protein-coding genes on chromosome 17 (see Methods), annotating 1,266 gene loci and 274 pseudogene loci (Supplementary Table S3). This catalogue includes all of the 1,079 genes on the chromosome in the RefSeq database (see Methods and Supplementary Information). The gene density (16.2 genes per Mb) is substantially higher than the genome average² and ranks only behind chromosome 19 (26.2 genes per Mb; ref. 3). There is evidence of extensive alternative splicing, with gene loci having an average of five distinct transcripts, and 76.6% having at least two transcripts, a proportion comparable to recent reports^{3,16–19}. Of the 274 pseudogenes on the chromosome, 73.0% are processed

(intronless) pseudogenes arising from retroposition. In addition to protein-coding genes, we identified 42 transfer RNA genes on the chromosome (Supplementary Table S4).

As part of the Mouse Genome Sequencing Consortium (MGSC), a finished sequence of mouse chromosome 11 has been produced (Supplementary Table S5). The sequence was derived from the C57BL/6J strain and has a finished euchromatic length of 118,816,080 bases, representing ~4.6% of the mouse genome. The finished sequence is interrupted by only two euchromatic gaps, with an estimated total size of 97 kb, and one additional gap containing the acrocentric heterochromatin (Fig. 1, Supplementary Table S1 and Methods). See Supplementary Table S5 for a clone list of the most up-to-date version of mouse chromosome 11.

Mouse chromosome 11 can be divided into two parts on the basis of conserved synteny with human chromosomes (Fig. 1). Although the proximal 59 Mb shows conserved synteny with five different human chromosomes (1, 2, 5, 7 and 22), the distal ~60 Mb has conserved synteny exclusively with human chromosome 17 (see Fig. 1 and Supplementary Fig. S1). The distal portion of mouse chromosome 11 is similar to human chromosome 17 in various respects, including having a high G+C content (45.5%), high gene density (18.7 genes per Mb), an excess of SINEs (14.4%), and a deficit of LINEs (8.2%). The ratio of SINEs to LINEs is similar to that in human, although the absolute density is lower, consistent with the fraction of detectable transposon sequence in mouse. In contrast, the proximal portion of mouse chromosome 11 has a lower G+C content (41.9%), low gene density (8.2 genes per Mb), and an opposite ratio of SINEs and LINEs (7.8% versus 19.1%). Although it is similar to human chromosome 17 in most other respects, the distal region of mouse chromosome 11 differs sharply with respect to segmental duplications: such sequences constitute only 1.1% of the distal region. Indeed, mouse chromosome 11 has the lowest rate of segmental duplication (1.4%) among all mouse chromosomes.

The block of undirected synteny (corresponding regions of orthology between the species, in which order may be disrupted by internal rearrangements) between human chromosome 17 and mouse chromosome 11 is the second largest such autosomal block between human and mouse (see Supplementary Information). Within this block, there are 23 segments of directed (collinear) synteny of >100 kb, indicating frequent inversion. To better understand this relationship, we reconstructed the chromosomal structure of the primate–rodent ancestor of human chromosome 17 (Supplementary Fig. S2) on the basis of genome sequences from human, mouse, dog and opossum (see Supplementary Information). The primate–rodent ancestor could be resolved except for one small segment, and also represents the state of the boreoeutherian ancestor except in two regions of conserved directed synteny between human and mouse that differ from dog and opossum (see Supplementary Information and Supplementary Fig. S2).

Comparison of the modern chromosomes with our ancestral reconstruction (Fig. 2a) shows that the mouse structure is virtually unchanged, whereas significant rearrangements have occurred in human. There are 20 rearrangement breakpoints in human but only three in mouse. (The terminal break and one internal break are used for different events in both species, leading to only 22 observed differences in the map of direct human–mouse conserved synteny; see Supplementary Fig. S1.) This is contrary to the general pattern for the human and mouse genomes, which typically shows many more rearrangements along the rodent lineage^{5,6} (see Supplementary Information). The mouse-specific rearrangements are confined to a single region orthologous to the Smith–Magenis region in human, which shows an even more complex pattern of rearrangements at finer scale (Fig. 2b).

What accounts for the extreme rearrangement of human chromosome 17? The answer seems to be related to the extensive segmental duplication on the human chromosome. The human-specific breaks

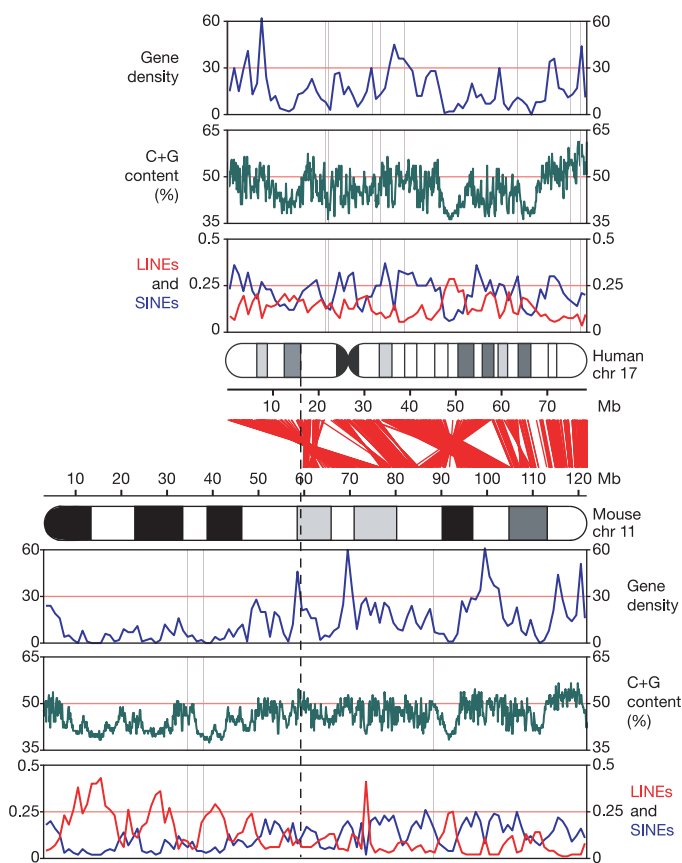


Figure 1 | Landscapes of human chromosome 17 and mouse chromosome 11. Approximate alignments of ideograms of the two chromosomes are shown in the centre of the figure, with red lines showing the relationships between orthologous genes. From top to bottom for each organism, tracks show gene density (in genes per Mb), G+C content on a scale from 35–65%, and densities of LINEs (red) and SINEs (blue) (fraction of bases). The vertical dashed line represents the approximate boundary between the distal region of mouse chromosome 11, which has shared synteny with human chromosome 17, and the proximal region, which does not. Chr, chromosome.

in conserved synteny are highly correlated with regions of intrachromosomal duplication (Fig. 3). Roughly 74% of the duplicated bases reside in the breakpoint regions between stretches of conserved synteny, which comprise only 7.3% of the bases in the chromosome sequence.

To further explore the nature of these duplicated sequences, we clustered the duplicons into classes on the basis of shared sequence (see Methods). After clustering, most of the duplicated sequences fall into one of three classes (1–3), which account for 51.6%, 19.9% and 3.1%, respectively, of the intrachromosomal duplications (Fig. 3 and Table 1). For each duplication class, we then sought to define a 'core element', corresponding to the maximum length of sequence occurring in the largest number of duplicons (see Methods). These core elements can serve as useful markers for identifying class members and may provide insight into the origins of the duplications²⁰.

The class 1 core element (~11 kb in length) contains a TBC1 domain and is found in 12 of the 19 human-specific breakpoint regions. Several of the class 1 core elements are actively transcribed, including six very recently duplicated genes from the TBC1D3 family (at 31–33 Mb) and the *USP6* oncogene (also known as *TRE-2*), which is a fusion of a TBC1D3 gene and the *USP32* gene²¹. The TBC1D3 family genes encode proteins with TBC1 domains, and are believed to be GTPase-activating proteins specific to *RAB5*, a RAS-related gene involved in vesicle trafficking. The TBC1D3 gene family has no known homologues outside the primates, and we could find no similar sequence in dog or mouse, making it impossible to identify an ancestral copy. However, the element certainly pre-dates the primate clade: human copies show substantial divergence (~34%) (Supplementary Fig. S3) and copies can be identified in ring-tailed lemur by polymerase chain reaction (PCR) (see Methods). The element is also duplicated in macaque, and a phylogeny of the elements in human and macaque (Supplementary Fig. S3) shows that at least some of these duplications pre-date the divergence of Old World monkeys, as some macaque copies have a human copy as their nearest neighbour.

The class 2 core element (~4 kb in length) is found in 9 of the 19

human-specific breakpoint regions (Fig. 3). The ancestral copy is probably the copy immediately upstream of the *NSF* gene at 41.95 Mb, as this is the only copy with an orthologous sequence in mouse. The class 2 duplicons occur primarily in three regions of chromosome 17q, at 26–27 Mb, 40–42 Mb and 60–63 Mb. These regions are all sites of extensive, small local inversion events that have occurred since the primate–rodent ancestor. A phylogenetic tree of the core elements (Supplementary Fig. S4) shows two main clusters, one for the copies at 26–27 Mb and another for those at 40–42 and 60–63 Mb. Notably, the reconstruction of the ancestral chromosome order (Supplementary Fig. S2) places these latter regions nearly adjacent to one another, suggesting that the core elements dispersed in a local region that was then disrupted by an inversion with breakpoints at 33 and 57 Mb. This ancestral chromosomal order seems to correspond to the modern karyotype of orangutan²², which has a paracentric inversion with respect to the human karyotype on 17q, with nearly coincident breakpoints. This would imply that the inversion in the human lineage is a fairly recent event. Notably, two class 2 core elements (the ancestral copy and a copy at 40.97 Mb) flank the boundaries of a 900-kb polymorphic inversion in the human population²³ and also the boundaries of a human rearrangement with respect to the ancestral chromosome (Fig. 2 and Supplementary Fig. S2), indicating a reuse of these breakpoints within the human lineage (see Supplementary Information).

The class 3 duplicons all reside in the CMT1A/SMS region (13.8–20.4 Mb), which seems to be a region of remarkable genomic fragility. The class is too small to define a meaningful core element, but examination of the longest copy shows that it is a previously identified low-copy repeat, LCRA1¹⁰. This element is the result of a fusion of two distinct 3'-UTR elements from elsewhere in the human genome. Although at least one of the individual sub-elements comprising the fusion is found in mouse and dog, the fusion product is not seen in these species, but only in primates. One of the two sub-elements is actively transcribed (see Supplementary Information).

Previous studies^{8,10} of the CMT1A/SMS region have identified large, highly similar local repeat blocks. The blocks are themselves

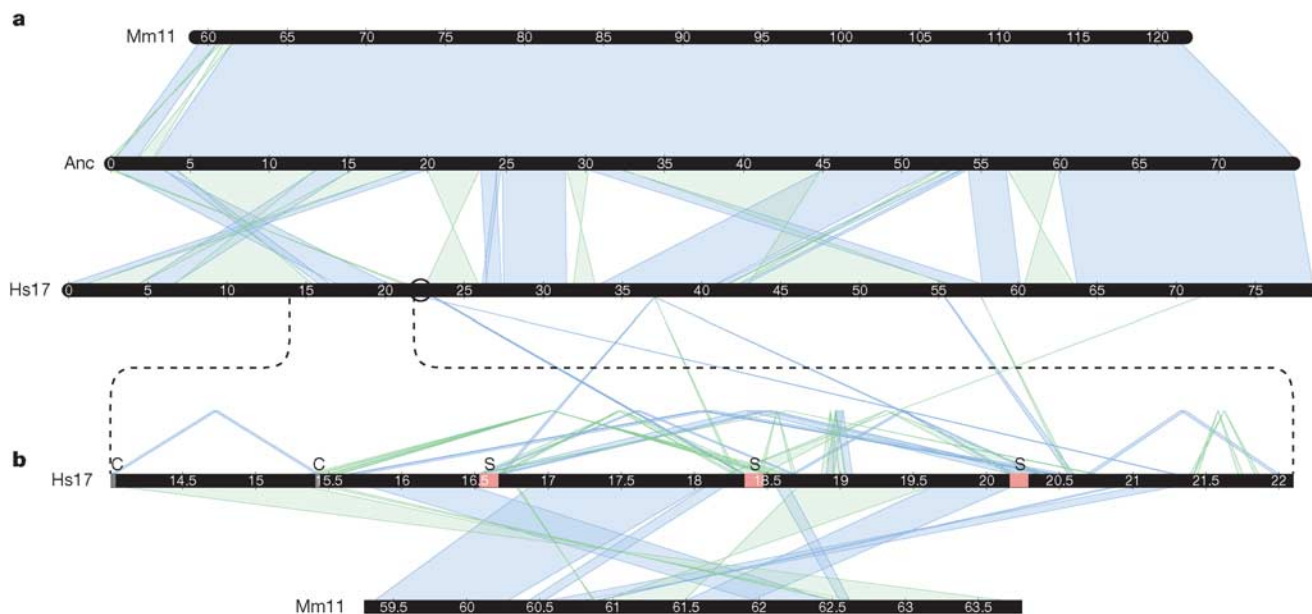


Figure 2 | Syntenic relationship between mouse, human and the ancestral chromosome. **a**, Conserved syntenic blocks of 100 kb or more between human chromosome 17 (Hs17), mouse chromosome 11 (Mm11), and a most parsimonious primate–rodent ancestor (Anc) reconstructed using dog and opossum as outgroups. **b**, An enlargement of the CMT1A/SMS region and the mouse orthologous sequence (human 14–22 Mb, mouse 59–64 Mb), including conserved syntenic blocks of 10 kb or more between human and mouse, and human segmental duplications of 20 kb or more. Segmental

duplications are shown above the human CMT1A/SMS line. Duplications with copies outside the enlarged region connect to 17q in **a**. Coloured blocks marked 'C' represent CMT1A duplicons, and those marked 'S' represent SMS duplicons. In both sections, direct (reference strand to reference strand) blocks of conserved synteny or segmental duplication are shown in blue; inverted blocks are shown in green. Chromosome blocks are labelled in megabases. The black circle indicates the location of the human chromosome 17 centromere.

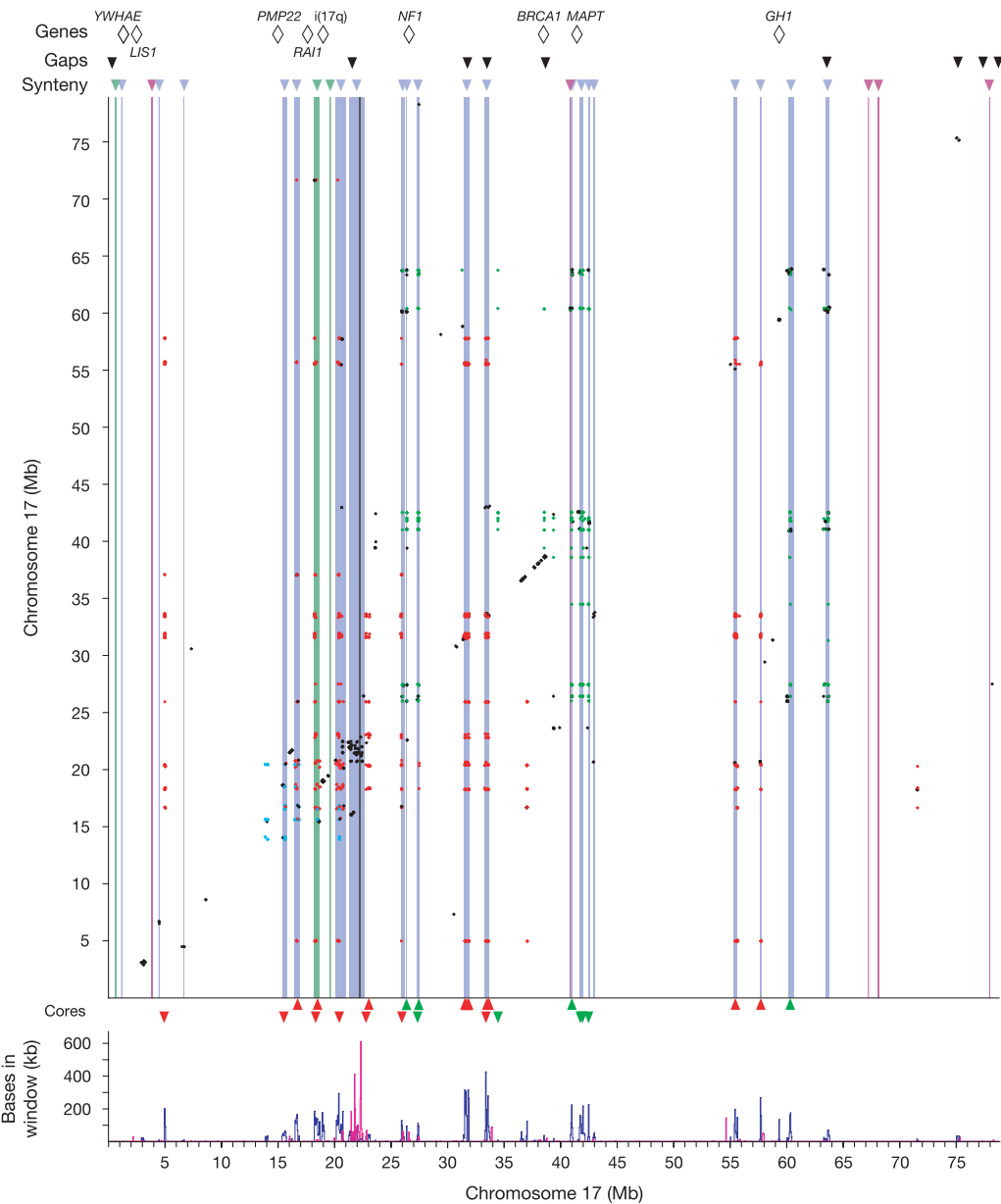


Figure 3 | Duplication landscape of chromosome 17 and its association with breaks in conserved synteny. The top line shows genes/breakpoints of interest for rearrangements mentioned in the text. Black triangles show the locations of the nine remaining gaps in the build. Below that, coloured triangles show breaks in synteny (blue for human, green for mouse, pink for dog). Intrachromosomal duplications are shown as an all-versus-all dot plot, coloured by class (class 1 in red, class 2 in green, class 3 in blue, all other classes in black; see Methods). Breaks in conserved synteny are represented as vertical bands coloured as above. The centromere is shown as a vertical dark grey band. Triangles at the bottom show core element locations (class 1 core elements in red, class 2 core elements in green), with direction showing orientation (up, forward strand; down, reverse strand). The panel at the bottom shows a density plot of intrachromosomal (blue) and interchromosomal (red) duplication, using 50-kb non-overlapping sliding windows.

fusions of multiple duplcons (including class 1 core elements) that do not occur together elsewhere in the genome. None of these duplicated sequences has a mouse orthologue within the region (Fig. 2b), having all come from the q arm of 17, indicating that this region may be a site of multiple inversion. Indeed, cytogenetic mapping of primate rearrangements²² shows the CMT1A/SMS region to be the site of an inversion breakpoint in orangutan and of a translocation between chromosomes 17 and 5 in gorilla²⁴. Certain common characteristics emerge from the study of segmental duplication on chromosome 17, as well as from our recent study of segmental duplications on chromosome 15 (ref. 20). First,

the vast majority of duplcons in these classes fall in breaks of conserved synteny. Second, segmentally duplicated sequence seems to spread readily within local regions. Local clustering is evident for all duplcon classes studied on chromosomes 15 and 17. Third, segmentally duplicated sequence then seems to spread over larger distances by chromosomal inversions. In two cases (class 2 on chromosome 17 and the main class on chromosome 15; ref. 20), such dispersal is confirmed on the basis of both sequence phylogeny and reconstruction of ancestral chromosome structure. With additional comparative data on lower primates, it should be possible to reconstruct the dispersal of most of the segmentally duplicated

Table 1 | Duplication classes on human chromosome 17

Class	No. of duplications	No. of duplications containing core element	No. of duplications in syntenic break regions	Core element size (bp)	Total bases in duplcons	Duplcon bases in syntenic break regions	Total duplcon bases in syntenic break regions (%)
Class 1	529	225	509	10,734	1,837,890	1,581,557	86.05
Class 2	204	65	202	4,185	979,083	735,413	75.11
Class 3	32	0	30	NA	125,758	105,741	84.08
Other (132)*	260	0	196	NA	2,449,945	1,574,952	64.29

NA, not applicable.
* Each of these 132 duplcon classes has fewer than 15 events.

sequences in the human genome. Finally, the core elements underlying the main classes of segmentally duplicated sequence on chromosomes 15 and 17 all show active transcription from a large number of the duplicons^{10,25,26}. This observation suggests a link between active transcription, duplication and rearrangement. It is possible that the nature of transcription, the chromatin structure or the transcripts at these loci may render the regions more susceptible to breakage, duplication or non-allelic recombination.

Our comparison of human chromosome 17 and distal mouse chromosome 11 shows that a genomic region can undergo strikingly different rearrangements in different lineages. There has been remarkably little interchromosomal rearrangement in the region orthologous to human chromosome 17 across all mammals, which may perhaps reflect an inherent stability of the region or a selective constraint on contiguity. In contrast, the region shows extensive internal rearrangement in the human lineage (20 breakpoints relative to the ancestral genome) but little rearrangement in mouse (only three rearrangements relative to the ancestor). The occurrence of the human breakpoints near segmental duplications supports a mechanistic link between these two features that may explain the unusual history of this genomic region in the primate lineage. Together with our recent analysis of chromosome 15, this analysis sheds light on the forces that shape the large-scale structure of genomes and, in the process, create both evolutionary variation and the chromosomal fragility underlying human disease.

METHODS

The finished sequence of human chromosome 17 was generated almost entirely (>99%) at the Broad Institute of MIT and Harvard (formerly the Whitehead Institute/MIT Center for Genome Research) (Supplementary Table S6). Chromosomal positions in this paper refer to NCBI human build 35. The finished sequence of mouse chromosome 11 was generated almost entirely (>99%) at the Sanger Institute (Supplementary Table S6). Chromosomal positions in this paper refer to NCBI mouse build 34. Methods for clone mapping, sequencing and tiling-path validation are described in the Supplementary Information. The gene catalogue and annotation for human chromosome 17 were produced as previously published¹⁶.

Syntenic maps and segmental duplications. Syntenic maps and segmental duplications were constructed as previously described^{15,16,20}.

Duplication class clustering. Pairwise duplications were clustered by physical overlap of ≥ 150 bp and at least 5% of the smaller element. We extended this by transitive closure to build maximally linked sets (see Supplementary Information).

Identification of core elements of duplicons. For each duplicon class, aligned coverage of each base in the duplicated region (the union of all duplicons) was recorded. Core elements were chosen by visual inspection to identify the longest contiguous sequence in the most deeply covered region. For each duplicon class, the longest core element sequence was aligned to the entire genome with PatternHunter²⁷, and all full-length alignments were identified as core element copies.

Construction of core element phylogeny. Construction of core element phylogeny was performed as previously described²⁰.

Received 9 February; accepted 1 March 2006.

1. Debry, R. W. & Seldin, M. F. Human/mouse homology relationships. *Genomics* **33**, 337–351 (1996).
2. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
3. Grimwood, J. *et al.* The DNA sequence and biology of human chromosome 19. *Nature* **428**, 529–535 (2004).
4. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
5. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).
6. Bourque, G., Zdobnov, E. M., Bork, P., Pevzner, P. A. & Tesler, G. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res.* **15**, 98–110 (2005).
7. Chance, P. F. *et al.* DNA deletion associated with hereditary neuropathy with liability to pressure palsies. *Cell* **72**, 143–151 (1993).

8. Park, S. S. *et al.* Structure and evolution of the Smith–Magenis syndrome repeat gene clusters, SMS-REPs. *Genome Res.* **12**, 729–738 (2002).
9. Chen, K. S. *et al.* Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome. *Nature Genet.* **17**, 154–163 (1997).
10. Inoue, K. *et al.* The 1.4-Mb CMT1A duplication/HNPP deletion genomic region reveals unique genome architectural features and provides insights into the recent evolution of new genes. *Genome Res.* **11**, 1018–1033 (2001).
11. Lupski, J. R. *et al.* DNA duplication associated with Charcot–Marie–Tooth disease type 1A. *Cell* **66**, 219–232 (1991).
12. Lupski, J. R. & Garcia, C. A. in *The Metabolic and Molecular Basis of Inherited Diseases* (eds Scriver, C. R., Beaudet, A. L., Sly, W. S. & Valle, D.) 5759–5788 (McGraw-Hill, New York, 2001).
13. Potocki, L. *et al.* Molecular mechanism for duplication. 17p11.2—the homologous recombination reciprocal of the Smith–Magenis microdeletion. *Nature Genet.* **24**, 84–87 (2000).
14. Barbouti, A. *et al.* The breakpoint region of the most common isochromosome, i(17q), in human neoplasia is characterized by a complex genomic architecture with large, palindromic, low-copy repeats. *Am. J. Hum. Genet.* **74**, 1–10 (2004).
15. Bailey, J. A. *et al.* Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001).
16. Nusbaum, C. *et al.* DNA sequence and analysis of human chromosome 18. *Nature* **437**, 551–555 (2005).
17. Nusbaum, C. *et al.* DNA sequence and analysis of human chromosome 8. *Nature* **439**, 331–335 (2006).
18. Hillier, L. W. *et al.* Generation and annotation of the DNA sequences of human chromosomes 2 and 4. *Nature* **434**, 724–731 (2005).
19. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 10. *Nature* **429**, 375–381 (2004).
20. Zody, M. C. *et al.* Analysis of the DNA sequence and duplication history of human chromosome 15. *Nature* **440**, 671–675 (2006).
21. Paulding, C. A., Ruvolo, M. & Haber, D. A. The *Tre2* (*USP6*) oncogene is a hominoid-specific gene. *Proc. Natl Acad. Sci. USA* **100**, 2507–2511 (2003).
22. Yunis, J. J. & Prakash, O. The origin of man: a chromosomal pictorial legacy. *Science* **215**, 1525–1530 (1982).
23. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nature Genet.* **37**, 129–137 (2005).
24. Stankiewicz, P., Park, S. S., Inoue, K. & Lupski, J. R. The evolutionary chromosome translocation 4;19 in *Gorilla gorilla* is associated with microduplication of the chromosome fragment syntenic to sequences surrounding the human proximal CMT1A-REP. *Genome Res.* **11**, 1205–1210 (2001).
25. Eichler, E. E. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* **17**, 661–669 (2001).
26. Stankiewicz, P., Shaw, C. J., Withers, M., Inoue, K. & Lupski, J. R. Serial segmental duplications during primate evolution result in complex human genome architecture. *Genome Res.* **14**, 2209–2220 (2004).
27. Ma, B., Tromp, J. & Li, M. PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18**, 440–445 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Special thanks are due to L. Gaffney for help with figures and tables. We thank T. Furey for help with lists of genetic markers and placement of RefSeqs, and K. Lindblad-Toh for sharing data on the opossum genome. We thank A. Kong for providing updated genetic map data, T. Hudson for technical advice, and the members of the Baylor College of Medicine Human Genome Sequencing Center, the J. Craig Venter Institute Joint Technology Center, and the Washington University Genome Sequencing Center for generation and early release of the genome assembly of the rhesus macaque. We also acknowledge the HUGO Gene Nomenclature Committee (S. Povey (chair), E. A. Bruford, R. C. Lovering, M. J. Lush, K. M. B. Sneddon, T. P. Sneddon, C. C. Talbot Jr and M. W. Wright) for assigning official gene symbols for human chromosome 17. We are grateful to all the members, present and past, of the Broad (and Whitehead) and Wellcome Trust Sanger sequencing platforms for their dedication and the consistent high quality of their data. The sequencing of human chromosome 17 was supported by a grant to the Whitehead Institute Center for Genome Research (now the Broad Institute) from the National Human Genome Research Institute (NHGRI). The sequencing of mouse chromosome 11 was supported by a grant to the Sanger Institute by the Wellcome Trust.

Author Information Accession numbers for all clones contributing to the finished sequence of human chromosome 17 can be found in Supplementary Table S2, and for mouse chromosome 11 in Supplementary Table S5. The updated human chromosome 17 sequence can be accessed through GenBank accession number NC_000017. The updated mouse chromosome 11 sequence can be accessed through the accession numbers listed in Supplementary Table S5. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.C.Z. (mczody@broad.mit.edu) or C.N. (chad@broad.mit.edu).

LETTERS

Repeated morphological evolution through *cis*-regulatory changes in a pleiotropic gene

Benjamin Prud'homme^{1*}, Nicolas Gompel^{1,2*}, Antonis Rokas^{1†}, Victoria A. Kassner¹, Thomas M. Williams¹, Shu-Dan Yeh³, John R. True³ & Sean B. Carroll¹

The independent evolution of morphological similarities is widespread^{1,2}. For simple traits, such as overall body colour, repeated transitions by means of mutations in the same gene may be common^{3–5}. However, for more complex traits, the possible genetic paths may be more numerous; the molecular mechanisms underlying their independent origins and the extent to which they are constrained to follow certain genetic paths are largely unknown. Here we show that a male wing pigmentation pattern involved in courtship display has been gained and lost multiple times in a *Drosophila* clade. Each of the cases we have analysed (two gains and two losses) involved regulatory changes at the pleiotropic pigmentation gene *yellow*. Losses involved the parallel inactivation of the same *cis*-regulatory element (CRE), with changes at a few nucleotides sufficient to account for the functional divergence of one element between two sibling species. Surprisingly, two independent gains of wing spots resulted from the co-option of distinct ancestral CREs. These results demonstrate how the functional diversification of the modular CREs of pleiotropic genes contributes to evolutionary novelty and the independent evolution of morphological similarities.

To address how complex traits are repeatedly gained and lost, we focused on the formation of spots of dark pigment at the tip of male wings in certain fruitfly (*Drosophila*) species of the *melanogaster*⁶ and *obscura*⁷ groups. These spots form under the control of multiple genes⁸. To determine how often male wing pigment spots have been gained or lost we established the phylogeny of these groups and reconstructed the evolution of the trait by means of bayesian phylogenetic inference (BI), which provides a statistical framework that explicitly accommodates phylogenetic and character mapping uncertainties in its ancestral character reconstruction estimates^{9–11}, and thus offers a more rigorous picture of character evolution.

The BI analysis provided evidence that the common ancestor of the clade was unspotted (Fig. 1, node 1) and that the wing spot was gained once within the *melanogaster* group (Fig. 1, node 2) and at least once more within the *obscura* group (Fig. 1, node 3). The analysis also indicated that the wing spot has been lost independently at least five times in the *melanogaster* group (Fig. 1, dashed branches). The availability of closely related species that differ with regard to wing spots enabled us to investigate the molecular mechanisms by which the spots were lost and gained.

D. elegans and *D. gunungcola* (Fig. 1, node 4) are interfertile sibling species that diverged 2–2.8 Myr ago and are characterized by differences in male-specific wing pigmentation¹² (Fig. 2b, c). The *yellow* (*y*) gene is a strong candidate for contributing to the difference in pigmentation because of its role in the production of black pigment in many *Drosophila* species^{8,13,14} including *D. elegans* (S.-D.Y. and

J.R.T., unpublished observations). We found that in *D. elegans* pupal wings Yellow accumulates at high levels in the region where the adult pigmentation spot will form (Fig. 2b, d), whereas in *D. gunungcola* the Yellow protein is only expressed at low levels throughout the wing, correlating with the grey shading of the adult wing (Fig. 2c, e).

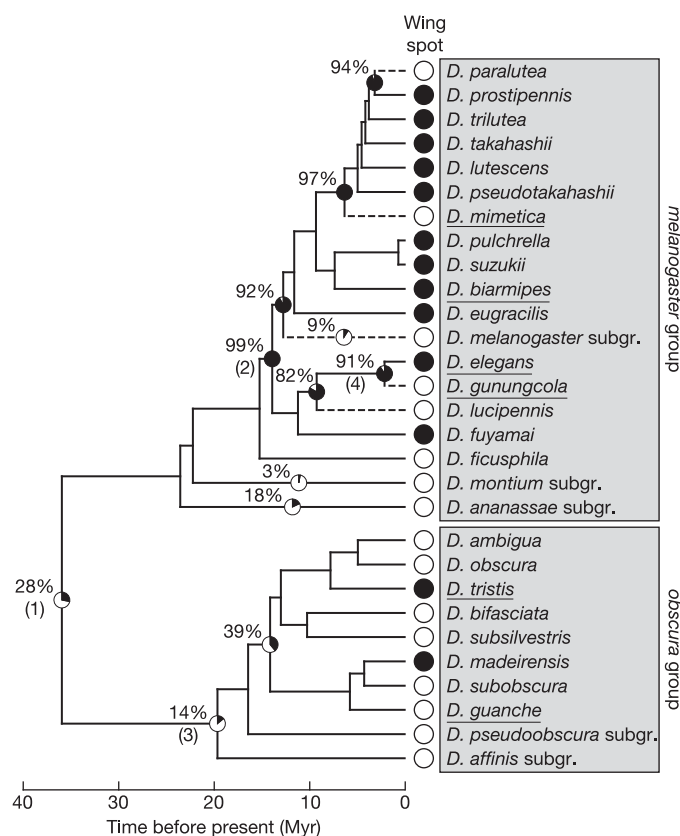


Figure 1 | Two independent gains and five losses of wing pigmentation spots in a *Drosophila* clade. For each species, the presence or absence of a wing spot is indicated with a solid black or a solid white circle, respectively. For key nodes, the black portion of each pie chart and the percentages shown next to them indicate the posterior probability that the ancestor was spotted. Dotted branches indicate inferred losses of the wing spot. Branch lengths correspond to absolute time estimates. Numbers in parenthesis identify nodes discussed in the text. Species examined in this study are underlined.

¹University of Wisconsin and Howard Hughes Medical Institute, Bock Laboratories, 1525 Linden Drive, Madison, Wisconsin 53706, USA. ²Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK. ³Department of Ecology and Evolution, 650 Life Sciences Building, Stony Brook University, Stony Brook, New York 11794, USA. [†]Present address: Broad Institute of MIT and Harvard, 320 Charles Street, Cambridge, Massachusetts 02141-2023, USA.

*These authors contributed equally to this work.

The divergence with regard to *yellow* function between the two species is therefore regulatory in nature.

As a first step to determining whether regulatory changes have occurred at γ in *D. gunungcola*, we sought to confirm that the formation of the wing spot of *D. elegans* and *D. biarmipes* share a common genetic basis. In *D. biarmipes*, the expression of γ in the future wing spot region is controlled by the *spot^{bia}* CRE, which evolved in the context of an ancestral CRE (*wing large*, Fig. 2a) driving uniform expression of γ in the wings of unspotted species⁸. We isolated from *D. elegans* a fragment orthologous to the *spot^{bia}* CRE, *spot^{ele}* (Fig. 2a), and assayed its transcriptional regulatory activity in *D. melanogaster*. This fragment drives reporter expression in a pattern specifically confined to the spot region (Fig. 2f). These results indicate that the wing pigmentation spots of *D. elegans* and *D. biarmipes* (and, by phylogenetic inference, all of the wing-spotted species of the *melanogaster* group descended from the ancestor at node 2 in Fig. 1) are homologous.

We then isolated from the *D. gunungcola* γ locus the 5' non-coding DNA and the fragment orthologous to the *spot* CRE (*spot^{gun}*). When transferred into *D. melanogaster* the 5' non-coding DNA drives uniform reporter expression in the wing, recapitulating the native *D. gunungcola* γ expression pattern (not shown). In contrast, the *spot^{gun}* CRE is largely inactive and drives barely detectable traces of reporter expression in the spot region (compare Fig. 2g with Fig. 2f). These results show that the major changes responsible for the loss of expression of *yellow* in *D. gunungcola* occurred in the *spot^{gun}* CRE.

The close kinship of *D. elegans* and *D. gunungcola* allowed us to pursue the molecular characterization of the functional divergence of the *spot* CRE. The *spot^{ele}* (775 base pairs (bp)) and *spot^{gun}* (724 bp) CREs share 92% sequence identity (Fig. 3a). Fifty nucleotide substitutions and five indels (four of less than 4 bp and one of 47 bp) constitute all of the differences between these two functionally divergent sequences. To map the sequence changes responsible for

the loss of activity of the element in *D. gunungcola* we constructed a series of *spot^{ele}* chimaeric transgenes (I to VI, Fig. 3b) in which we systematically replaced subsets of *D. elegans*-specific nucleotides with their divergent *D. gunungcola* counterparts and assayed their transcriptional regulatory activity in *D. melanogaster*. One obvious candidate for the loss of activity was the 47-bp indel present in the *spot^{ele}* sequence but absent from *spot^{gun}*. However, when this 47-bp stretch was selectively removed from the *spot^{ele}* element (construct Δ), the transgene drove reporter expression similar to the native *spot^{ele}* sequence (Fig. 3c). In fact, we found that most chimaeric constructs drove reporter levels and patterns in a similar manner to the *spot^{ele}* element (Figs 2f and 3d, and not shown), thus ruling out significant functional contribution of these divergent nucleotides to the loss of activity of *spot^{gun}*. However, one chimaeric transgene, III, yielded a reporter expression similar to that of the *spot^{gun}* element (Figs 2g and 3e). This construct included ten divergent nucleotides, indicating that the major functional differences between the two species' *spot* elements might be due to changes among these sites. To further delimit the number of nucleotides responsible for the divergence in gene expression, we made three sub-constructs (III.1, III.2 and III.3) containing just three, four and three nucleotide differences, respectively. Whereas III.3 yielded a normal spot pattern, lines carrying III.1 or III.2 showed a strong reduction in reporter expression (not shown). These results indicate that divergent nucleotides within both III.1 and III.2 are necessary for the function of *spot^{ele}* and were involved in the functional change between the two species' CREs.

To test whether these divergent nucleotides contributed to the loss of *Yellow* spot expression in *D. gunungcola*, we made a reciprocal construct in which we replaced the ten divergent nucleotides of the *spot^{gun}* element with their *D. elegans* counterparts. This construct, *spot^{gun} rescue*, drives a spot expression pattern similar to that of *spot^{ele}*, thus restoring the ancestral function of the *spot^{gun}* regulatory element (Fig. 3b, f). This result demonstrates the critical contribution of these ten nucleotides to the divergence in transcriptional regulatory activity between the two species' CREs and indicates that the *spot^{gun}* CRE, although inactive, still contains some of the regulatory information required for generating the spot pattern. We deduce that at least two and no more than seven point mutations (presumably affecting two or more transcription-factor-binding sites) were sufficient to alter the function of the *spot^{gun}* element and fully account for the divergence in γ expression in the wing between *D. elegans* and *D. gunungcola*. We suggest that functional inactivation of CREs might require only a small number of mutational steps and might be the most likely path to the evolutionary loss of gene expression.

To test the generality of CRE inactivation in the loss of wing spots, we examined the expression and regulation of γ in *D. mimetica* (Fig. 1, and Supplementary Fig. 2). As in *D. gunungcola*, the spot loss in *D. mimetica* is associated with changes in *Yellow* expression in the wing. Furthermore, the *D. mimetica* *spot* CRE, which shares 69% sequence identity with *spot^{ele}*, drives a very faint reporter expression in the spot region, similar in intensity to that of the *spot^{gun}* CRE (Supplementary Fig. 2c). Taken together, our results show that the independent inactivation of the same CRE has contributed to the repeated loss of a wing pigmentation pattern in at least two species.

To determine whether the repeated gains of wing spots also had a common mechanistic basis, we examined *Yellow* expression in and CREs from *D. tristis*, which gained a wing spot independently from the *melanogaster* group species (Figs 1 and 4a). Remarkably, the trait has apparently evolved under sexual selection in both cases, along with a novel wing display step during the male courtship (Supplementary movies). *Yellow* expression in *D. tristis* male pupal wings, just as in spotted *melanogaster* group species, prefigures the wing pigmentation spot (Fig. 4a, b). We tested whether in *D. tristis* this expression was also controlled by the *wing large* CRE we had identified in *D. melanogaster* group species⁸. However, when

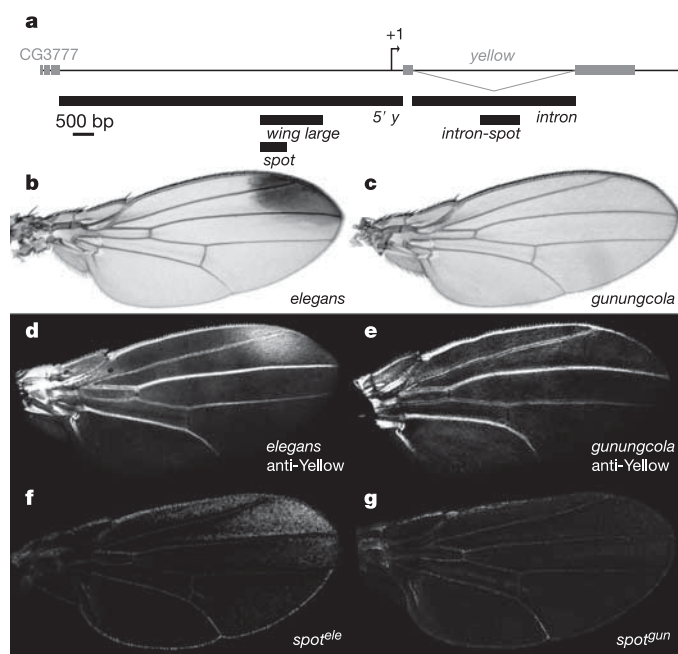


Figure 2 | Changes in the *yellow* spot cis-regulatory DNA underlie the loss of the pigmentation spot in *D. gunungcola*. **a**, Diagram of the *yellow* locus showing the position of the DNA fragments tested in this study (black bars). **b–e**, The adult male wings of *D. elegans* (**b**) and *D. gunungcola* (**c**) differ in their pigmentation patterns, which reflect the expression of the *Yellow* protein in developing pupal wings (**d** and **e**, respectively). **f, g**, The *D. elegans* CRE *spot^{ele}*, nested in the 5' region, drives high levels of reporter expression in the spot region (**f**), whereas its *D. gunungcola* orthologue, *spot^{gun}*, has lost this activity (**g**).

transferred into *D. melanogaster* this CRE drove only uniform reporter expression in the wing (Fig. 4c), as expected for the ancestral *wing large* element, but no elevated expression in the spot region was observed.

We then considered whether a distinct CRE controlling Yellow expression in the spot had evolved at another position of the locus. We found that none of the sequences extending over the 5' non-coding region of the locus drove any wing spot pattern (data not shown). However, the intron of the *D. tristis* *yellow* gene drove reporter expression in a spot pattern, as well as strong expression in the wing veins (Fig. 4d). Further dissection of the *D. tristis* intron revealed that expression in both the wing veins and the spot are controlled by a smaller fragment, *intron-spot* (927 bp; Fig. 2a, and data not shown). Because *D. tristis* contains a functional element orthologous to the *wing large* element of other species, and because we did not detect any sequence similarity between the *spot* and the *intron-spot* elements, we conclude that the evolution of *y* expression in the wing spot of *D. tristis* involved an entirely different *cis*-regulatory region. To determine whether the *D. tristis* *intron-spot* CRE evolved *de novo* or through the modification of a pre-existing element, we examined the transcriptional regulatory activity of the *y* intron from *D. guanche*, an unspotted species of the *obscura* group (Supplementary Fig. 3a). This sequence drove a wing vein expression pattern similar to that driven by the *D. tristis* *y* *intron-spot* CRE (Supplementary Fig. 3b). Thus, the proximity or overlap of an ancestral *wing vein* CRE and the derived *spot* CRE in the *D. tristis* *y* intron indicates that the spot activity might have evolved through the co-option of an ancestral *wing vein* CRE.

The molecular bases of spot evolution offer some general insights into how the course of evolution is determined by which genetic

paths are possible, which are most probable, and which are permissible under natural selection. The gains of *yellow* expression in the wing spots of the *melanogaster* group species and in *D. tristis* through the co-option and modification of distinct, ancestral CREs of the *yellow* locus show that multiple molecular paths were possible for the evolution of spot formation (Fig. 4e). Yet both ancestral elements drove *y* expression in the pupal wing. We have suggested that novel CREs active in a particular tissue are most likely to evolve in the vicinity of pre-existing elements active in that tissue⁸. Indeed, rather than evolving the constellation of binding sites required for element function *de novo*, it is more likely that a new activity would evolve in the context of a functional element that already contained information (for example, particular binding sites) for the control of a gene. The co-option of available transcription factors and the co-option and modification of CREs to generate novel patterns illustrate, at the level of regulatory circuits, the sort of 'evolutionary tinkering' envisaged by François Jacob three decades ago¹⁵.

Such tinkering is also a matter of what is permissible. Regulatory changes in one CRE, by selectively affecting one aspect of the spatio-temporal pattern of expression, circumvent the fitness reduction potentially associated with the global effects of changes in protein function. The *yellow* gene is highly pleiotropic^{16,17} and the control of the different Yellow expression domains is governed by distinct CREs. Other pleiotropic genes also seem to have contributed to repeated evolutionary transitions through CREs^{18–20}. Evolutionary changes in protein coding sequences have been observed in pigmentation genes in vertebrates, such as *MC1R*^{4,5} and *Oca3*, but these encode specialized, minimally pleiotropic proteins required for pigment production in melanocytes. These findings indicate that pleiotropy might be an important genetic constraint on the potential

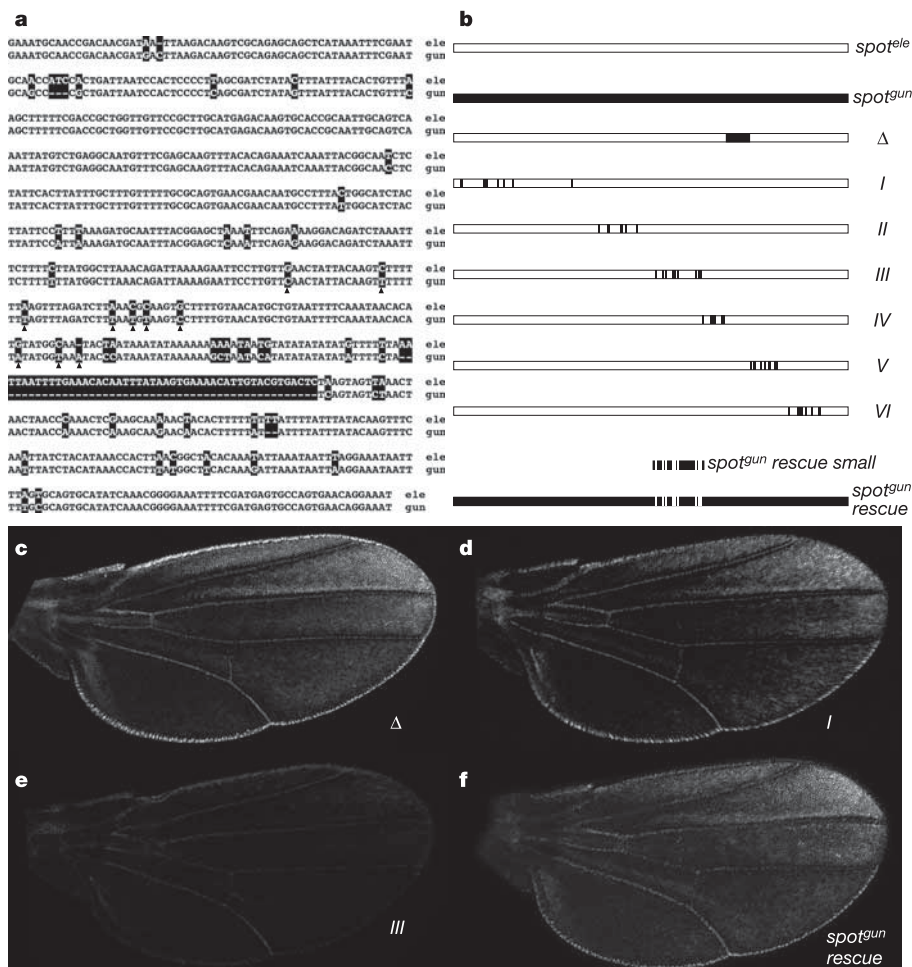


Figure 3 | A few divergent nucleotides account for the functional inactivation of the *spot*^{gun} element. **a**, *spot*^{ele} and *spot*^{gun} alignment; arrowheads indicate divergent nucleotides of construct III. **b**, Diagram of chimaeric constructs to map functional changes. **c–f**, Reporter expression in transformed pupal wings. Construct Δ (**c**) and I (**d**) do not affect spot expression. However, construct III (**e**) drives a dramatically reduced reporter expression, similar to that of *spot*^{gun} (Fig. 2g). The reciprocal construct to III, *spot*^{gun} rescue (**f**), restores a spot pattern. However, the 90-bp subfragment, *spot*^{gun} rescue small, containing the ten key divergent nucleotides, is not active (not shown).

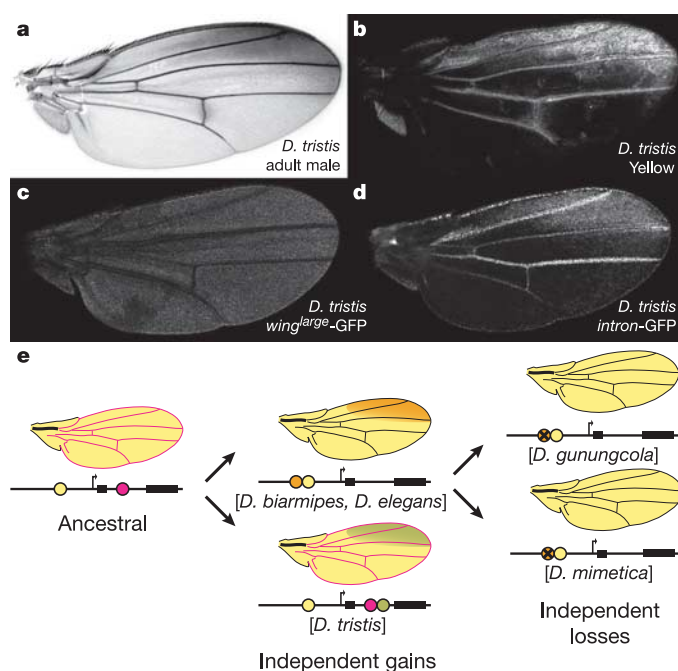


Figure 4 | Independent co-option of CREs of the yellow gene in wing spot evolution. **a, b,** In the *D. tristis* male pupal wing (**a**), Yellow expression prefigures the pigmentation pattern (**b**). **c, d,** The *D. tristis* *y* wing large orthologue (**c**) drives a uniform low level of reporter expression in the pupal wing, whereas the *D. tristis* *y* intron (**d**) drives the novel expression pattern in the spot region as well as the wing veins. **e,** *y* spot expression patterns evolved twice by the co-option and modification of two different pre-existing *y* CREs (symbolized by yellow → orange circles and pink → green circles) and were lost by the repeated inactivation of one CRE.

contribution of the different parts of genes (coding versus non-coding sequences) to morphological evolution. In highly pleiotropic genes, *cis*-regulatory changes are more likely to be accommodated than coding changes and should be expected to be major contributors to morphological evolution.

METHODS

Phylogenetics. We assembled a data matrix of 11 genes from 77 species (including new sequences and previously available data; Supplementary Tables 1 and 2) with an average of six genes per species. Selection of species was random in respect to their state of the wing spot trait and balanced across clades. Phylogenies were estimated under three optimality criteria: bayesian inference, maximum parsimony and maximum likelihood, as implemented in MrBayes²¹, PAUP²² and PHYML²³, respectively. The resulting phylogeny is generally consistent with previous studies with smaller numbers of genes and taxa^{24–28}. Ancestral character reconstruction was performed with BayesMultistate⁹. Divergence date estimates for the whole clade were obtained with the penalized likelihood method as implemented in r8s, version 1.7 (ref. 29). Details of all phylogenetic analyses are given in Supplementary Information, and the data matrix and analyses are available from the authors on request.

Immunohistochemistry, cloning and imaging. These were performed as described previously⁸ (specific details are provided in Supplementary Information).

Received 4 November 2005; accepted 25 January 2006.

- Conway Morris, S. *Life's Solution. Inevitable Humans in a Lonely Universe* (Cambridge Univ. Press, Cambridge, 2005).
- West-Eberhard, M. J. *Developmental Plasticity and Evolution* (Oxford Univ. Press, Oxford, 2003).
- Protas, M. E. *et al.* Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. *Nature Genet.* **38**, 107–111 (2006).
- Nachman, M. W., Hoekstra, H. E. & D'Agostino, S. L. The genetic basis of adaptive melanism in pocket mice. *Proc. Natl Acad. Sci. USA* **100**, 5268–5273 (2003).
- Theron, E., Hawkins, K., Bermingham, E., Ricklefs, R. E. & Mundy, N. I. The

molecular basis of an avian plumage polymorphism in the wild: a melanocortin-1-receptor point mutation is perfectly associated with the melanistic plumage morph of the bananaquit, *Coereba flaveola*. *Curr. Biol.* **11**, 550–557 (2001).

- Bock, I. R. & Wheeler, M. R. in *Studies in Genetics* (ed. Wheeler, M. R.) 1–102 (University of Texas, Austin, 1972).
- Lakovaara, S. & Saura, A. in *The Genetics and Biology of Drosophila* Vol. 3b (eds Ashburner, M., Carson, H. L. & Thompson, J. N.) 1–59 (Academic, London, 1982).
- Gompel, N., Prud'homme, B., Wittkopp, P. J., Kassner, V. A. & Carroll, S. B. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in *Drosophila*. *Nature* **433**, 481–487 (2005).
- Pagel, M., Meade, A. & Barker, D. Bayesian estimation of ancestral character states on phylogenies. *Syst. Biol.* **53**, 673–684 (2004).
- Huelsenbeck, J. P., Rannala, B. & Masly, J. P. Accommodating phylogenetic uncertainty in evolutionary studies. *Science* **288**, 2349–2350 (2000).
- Ronquist, F. Bayesian inference of character evolution. *Trends Ecol. Evol.* **19**, 475–481 (2004).
- Yeh, S.-D., Liou, S.-R. & True, J. R. Genetics of divergence in male wing pigmentation and courtship behavior between *Drosophila elegans* and *D. gunungcola*. *Heredity* (in the press).
- Wittkopp, P. J., Vaccaro, K. & Carroll, S. B. Evolution of yellow gene regulation and pigmentation in *Drosophila*. *Curr. Biol.* **12**, 1547–1556 (2002).
- Wittkopp, P. J., True, J. R. & Carroll, S. B. Reciprocal functions of the *Drosophila* yellow and ebony proteins in the development and evolution of pigment patterns. *Development* **129**, 1849–1858 (2002).
- Jacob, F. Evolution and tinkering. *Science* **196**, 1161–1166 (1977).
- Drapeau, M. D., Cyran, S. A., Viering, M. M., Geyer, P. K. & Long, A. D. A cis-regulatory sequence within the yellow locus of *Drosophila melanogaster* required for normal male mating success. *Genetics* **172**, 1009–1030 (2006).
- Walter, M. F. *et al.* Temporal and spatial expression of the yellow gene in correlation with cuticle formation and dopa decarboxylase activity in *Drosophila* development. *Dev. Biol.* **147**, 32–45 (1991).
- Shapiro, M. D. *et al.* Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428**, 717–723 (2004).
- Sucena, E., Delon, I., Jones, I., Payre, F. & Stern, D. L. Regulatory evolution of *shavenbaby/ovo* underlies multiple cases of morphological parallelism. *Nature* **424**, 935–938 (2003).
- Sucena, E. & Stern, D. L. Divergence of larval morphology between *Drosophila sechellia* and its sibling species caused by cis-regulatory evolution of *ovo/shaven-baby*. *Proc. Natl Acad. Sci. USA* **97**, 4530–4534 (2000).
- Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).
- Swofford, D. L. PAUP*: *Phylogenetic Analysis Using Parsimony* (*and Other Methods) 4.0 Beta (Sinauer Associates, Sunderland, Massachusetts, 2002).
- Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704 (2003).
- O'Grady, P. M. & Kidwell, M. G. Phylogeny of the subgenus *Sophophora* (Diptera: Drosophilidae) based on combined analysis of nuclear and mitochondrial sequences. *Mol. Phylogenet. Evol.* **22**, 442–453 (2002).
- O'Grady, P. M. Reevaluation of phylogeny in the *Drosophila obscura* species group based on combined analysis of nucleotide sequences. *Mol. Phylogenet. Evol.* **12**, 124–139 (1999).
- Goto, S. G. & Kimura, M. T. Phylogenetic utility of mitochondrial COI and nuclear Gpdh genes in *Drosophila*. *Mol. Phylogenet. Evol.* **18**, 404–422 (2001).
- Kopp, A. & True, J. R. Phylogeny of the oriental *Drosophila melanogaster* species group: a multilocus reconstruction. *Syst. Biol.* **51**, 786–805 (2002).
- Schwaroch, V. Phylogeny of a paradigm lineage: the *Drosophila melanogaster* species group (Diptera: Drosophilidae). *Biol. J. Linn. Soc.* **76**, 21–37 (2002).
- Sanderson, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301–302 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Kopp for sharing unpublished data; J. J. Gao, S. Hagemann, M. T. Kimura, A. Kopp, V. Stamatakis and the *Drosophila* Tucson Stock Center for providing fly stocks; M. Toda for fly identification; M. Bate for help with behavioural observations; A. Meade and M. Pagel for advice on bayesian ancestral character reconstruction; and B. Williams and J. Yoder for comments on the manuscript. N.G. thanks P. Simpson for hosting him in her laboratory. N.G. was an EMBO and a Human Frontier long-term fellow. The project was supported by the Howard Hughes Medical Institute (S.B.C.).

Author Information The sequences described in this paper have been deposited at the EMBL nucleotide database under accession numbers AM181668 to AM181680. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.B.C. (sbcarroll@wisc.edu).

LETTERS

Synaptic scaling mediated by glial TNF- α

David Stellwagen¹ & Robert C. Malenka¹

Two general forms of synaptic plasticity that operate on different timescales are thought to contribute to the activity-dependent refinement of neural circuitry during development: (1) long-term potentiation (LTP) and long-term depression (LTD), which involve rapid adjustments in the strengths of individual synapses in response to specific patterns of correlated synaptic activity, and (2) homeostatic synaptic scaling, which entails uniform adjustments in the strength of all synapses on a cell in response to prolonged changes in the cell's electrical activity^{1,2}. Without homeostatic synaptic scaling, neural networks can become unstable and perform suboptimally¹⁻³. Although much is known about the mechanisms underlying LTP and LTD⁴, little is known about the mechanisms responsible for synaptic scaling except that such scaling is due, at least in part, to alterations in receptor content at synapses⁵⁻⁷. Here we show that synaptic scaling in response to prolonged blockade of activity is mediated by the pro-inflammatory cytokine tumour-necrosis factor- α (TNF- α). Using mixtures of wild-type and TNF- α -deficient neurons and glia, we also show that glia are the source of the TNF- α that is required for this form of synaptic scaling. We suggest that by modulating TNF- α levels, glia actively participate in the homeostatic activity-dependent regulation of synaptic connectivity.

TNF- α increases the cell-surface expression of AMPA receptors (α -amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors, or AMPARs) in cultured hippocampal neurons and influences synaptic strength in both cultures and hippocampal slices^{8,9}. These findings raise the possibility that TNF- α might have an important role in NMDA (*N*-methyl-D-aspartate) receptor-dependent LTP and LTD, which similarly involve regulation of synaptic AMPARs⁴. To investigate whether TNF- α has a role in these forms of synaptic plasticity, we incubated acute rat hippocampal slices in high, saturating levels of TNF- α . Consistent with its effects in cultured neurons, incubating slices with TNF- α caused an increase in the ratio of AMPAR- to NMDAR-mediated synaptic currents (Fig. 1a). However, this same pretreatment did not affect the magnitude of LTP, nor did TNF- α treatment affect LTD (Fig. 1b) or the agonist-induced endocytosis of AMPARs in dissociated cultures (data not shown), a culture model for LTD¹⁰.

To further examine the requirement of TNF- α in rapidly induced, long-lasting synaptic plasticity, we assayed LTP and LTD in hippocampal slices prepared from mice lacking TNF- α (*Tnf* knockout mice)¹¹. Both LTP and LTD were normal in the absence of TNF- α -mediated signalling (Fig. 1c). Finally, we examined hippocampal slices from knockout mice lacking both TNF- α receptors, TNFR1 (*Tnfrsf1a*) and TNFR2 (*Tnfrsf1b*) (ref. 12). Similar to the *Tnf* knockout mice, hippocampal slices prepared from these mice expressed LTP and LTD equivalent to that observed in slices from wild-type mice (Fig. 1d).

Although these data indicate that TNF- α is not a requisite mediator of rapid forms of long-term synaptic plasticity, it influences AMPAR and GABA_A (γ -aminobutyric acid) receptor trafficking in a manner that shares key features with the homeostatic scaling of

excitatory and inhibitory synapses in response to prolonged blockade of activity^{1,5-7,13}. If a soluble protein such as TNF- α has a critical role in synaptic scaling, it should accumulate in the extracellular medium of hippocampal cultures subject to chronic activity blockade. To test this prediction, we treated naive cultures with media collected from sister cultures treated with tetrodotoxin for 48 h (TTX, which blocks voltage-gated Na⁺ channels and action potential generation) and then measured cell-surface levels of AMPARs. Consistent with previous results^{5,6,8}, both acute treatment with TNF- α and chronic treatment with TTX substantially increased AMPAR surface expression (Fig. 2a).

Notably, acute treatment of cultures with conditioned medium from the TTX-treated cultures was similarly effective at increasing cell-surface AMPAR levels (Fig. 2b). To test whether the effects of the conditioned medium on AMPARs required TNF- α , we added a soluble form of the TNFR1 receptor (sTNFR), which scavenges TNF- α ^{8,9,14}, to the medium before applying it to naive cultures. In the presence of sTNFR, conditioned medium from TTX-treated cultures no longer increased surface expression of AMPARs (Fig. 2b). Acute treatment of cultures with TTX alone had no effect on cell-surface AMPAR levels ($n = 66$; data not shown); 24-h TTX treatment was also insufficient to increase surface AMPARs ($n = 60$), as was the conditioned media from 24-h-treated cultures ($n = 60$; data not shown). These results suggest that TNF- α release is not rapidly modulated by activity, but that chronic activity blockade causes release of sufficient amounts of TNF- α to acutely increase surface AMPAR expression.

To test whether the observed increases in surface AMPAR levels occur at synapses in a manner that influences synaptic strength, we recorded miniature excitatory postsynaptic currents (mEPSCs) from presumptive pyramidal cells. In agreement with published results^{5,13,15}, chronic activity blockade led to an increase in mEPSC amplitude (Fig. 2c). Consistent with our immunocytochemical results, acute application of conditioned medium from TTX-treated cultures to naive sister cultures also caused an increase in mEPSC amplitude (Fig. 2c), an effect that was prevented by pretreating the conditioned medium with sTNFR (Fig. 2c). On average, the frequency of mEPSCs also increased after acute treatment with conditioned medium (data not shown), consistent with the effects of acute applications of TNF- α ⁸. However, this change was temporary, as no chronic manipulation had a significant effect on mEPSC frequency. These results suggest that chronic activity blockade causes the release of soluble TNF- α that is necessary for the increases in surface AMPAR levels and synaptic strength caused by conditioned medium from TTX-treated cultures. Consistent with this hypothesis, an L929 cell death assay¹⁶ revealed that the level of TNF- α in conditioned media from activity-blockaded cultures was significantly higher than in media from control cultures (Fig. 2d).

To test directly whether TNF- α is required for homeostatic synaptic scaling in response to activity blockade, and is not an incidental byproduct of manipulating activity, we interfered with TNF- α signalling during activity blockade by adding sTNFR during

¹Nancy Pritzker Laboratory, Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Palo Alto, California 94304-5485, USA.

the second day of the two-day TTX treatment. This prevented any increase in mEPSC amplitude due to the activity blockade while, as expected, causing a modest decrease in the control mEPSC amplitude⁸ (Fig. 2e). In addition to increasing excitatory synaptic strength, chronic activity blockade and acute application of TNF- α also decrease inhibitory synaptic strength^{7,9}. To test whether TNF- α is required for the decrease in the amplitude of miniature inhibitory postsynaptic currents (mIPSCs) caused by activity blockade⁷, we again added sTNFR during the second day of TTX treatment. This prevented the decrease in mIPSC amplitude caused by TTX treatment alone (Fig. 2f).

Our results thus far provide strong evidence that TNF- α is required for the scaling of both excitatory and inhibitory synapses in response to activity blockade. However, it is possible that a related ligand that binds TNF- α receptors is the actual signalling molecule required for this form of synaptic scaling. To test this hypothesis, we examined synaptic scaling in *Tnf* knockout (*Tnf*^{-/-}) mice. Although robust synaptic scaling was observed in cultures from wild-type mice, an increase in the amplitudes of mEPSCs in response to chronic activity blockade did not occur in hippocampal cultures from *Tnf*^{-/-} mice (Fig. 3a, b).

Synaptic scaling is a bidirectional phenomenon in which excitatory synapses scale up in response to activity reduction, but scale down in response to increases in activity^{1,2,5,15,17}. Although our data indicate that scaling up of excitatory synapses requires TNF- α

signalling, scaling down could result either from a reduction in constitutively produced TNF- α , which actively maintains surface levels of AMPARs⁸, or by signals that regulate synapses inversely to TNF- α . To determine whether signals other than TNF- α are responsible for homeostatic weakening of excitatory synapses^{1,2,5,15}, we treated *Tnf*^{-/-} mouse cultures for two days with the GABA_A receptor antagonist picrotoxin (PTX, 50 μ M), because GABA_A receptor currents are largely inhibitory at the culture age we studied¹⁸. PTX-treated *Tnf*^{-/-} cultures showed a reduction in mEPSC amplitudes equivalent to that observed in PTX-treated, wild-type mouse cultures (Fig. 3c, d). These results suggest that synaptic strength is homeostatically regulated by at least two opposing signals, with TNF- α shifting neurons towards more excitation and less inhibition, and other signals (perhaps such as brain-derived neurotrophic factor¹⁹) pushing neurons towards less excitation and more inhibition.

Virtually all previous work on synaptic scaling has been done in dissociated cultures, raising the possibility that the phenomenon is an artefact of the culture system. To address this concern, we studied the effects of activity deprivation on hippocampal slice cultures, a preparation that expresses normal forms of LTP and LTD and seems to traffic AMPARs in a manner indistinguishable from that seen in acute slice preparations^{20–22}. Similar to dissociated cultures, blocking activity in cultured slices from wild-type mice for two days caused an increase in the amplitude of mEPSCs recorded from CA1 pyramidal

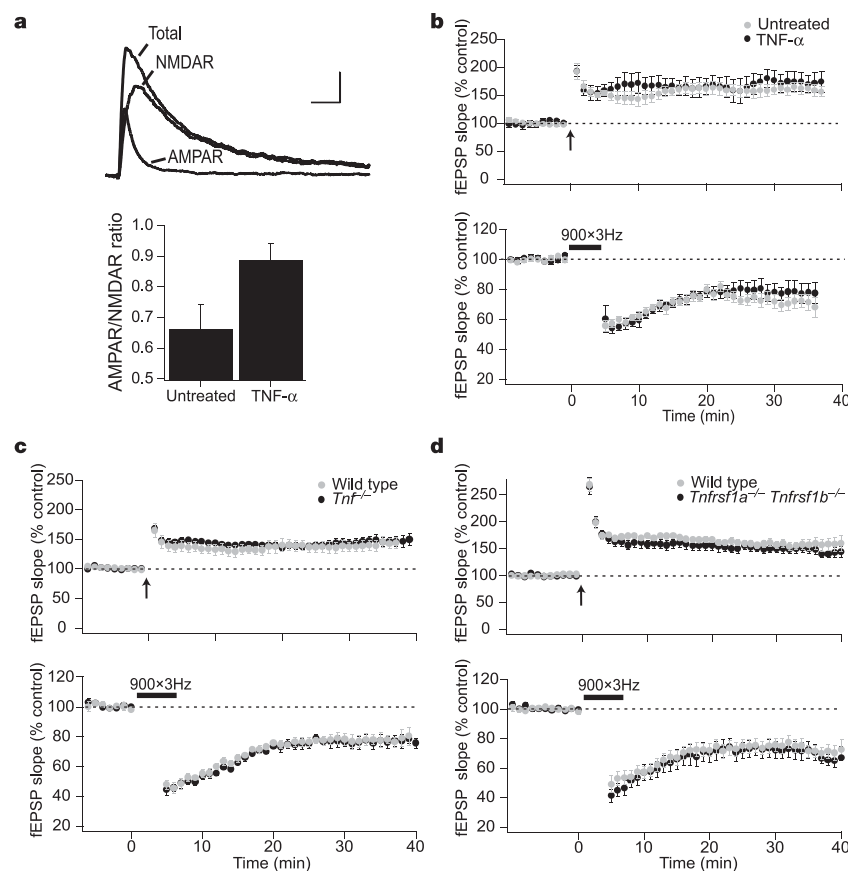


Figure 1 | TNF- α signalling is not required for CA1 hippocampal LTP or LTD. **a**, TNF- α treatment increases the ratio of AMPAR to NMDAR currents. Top, representative traces from whole-cell voltage-clamp recording at +40 mV, showing total current and isolated AMPAR and NMDAR currents (scale bars 25 pA, 50 ms). Bottom, group data showing that the AMPAR-to-NMDAR ratio is increased in cells from acute hippocampal slices treated with TNF- α (0.89 ± 0.05 , $n = 8$), compared with control slices from the same animals (0.66 ± 0.08 , $n = 7$, $P < 0.05$). **b**, Summary graphs showing LTP (top) (untreated, $162 \pm 10\%$, $n = 5$; TNF- α , $173 \pm 16\%$,

$n = 4$) and LTD (bottom) (untreated, $71 \pm 5\%$, $n = 8$; TNF- α , $78 \pm 6\%$, $n = 8$) of field EPSPs in slices treated with TNF- α or untreated slices from the same animals. **c**, LTP and LTD in hippocampal slices from *Tnf*^{-/-} mice (LTP, $139 \pm 9\%$, $n = 5$; LTD, $76 \pm 4\%$, $n = 6$) and wild-type mice (LTP, $145 \pm 8\%$, $n = 4$; LTD, $78 \pm 4\%$, $n = 3$). **d**, LTP and LTD in hippocampal slices from *Tnfrsf1a*^{-/-} *Tnfrsf1b*^{-/-} double knockout mice (LTP, $150 \pm 8\%$, $n = 9$; LTD, $72 \pm 6\%$, $n = 12$) and wild-type mice (LTP, $157 \pm 8\%$, $n = 11$; LTD, $75 \pm 4\%$, $n = 15$). Error bars show s.e.m.

cells (Fig. 3e, f). However, slices prepared from *Tnf*^{-/-} mice did not show synaptic scaling (Fig. 3e, f). These data suggest that synaptic scaling requiring TNF- α is a general phenomenon of neuronal circuits, and is not restricted to dissociated culture preparations. Consistent with this proposal, basal synaptic strength in acute hippocampal slices is influenced by constitutively released TNF- α ⁸.

Previous work has implied that neurons themselves sense their overall level of activity and generate the signals responsible for synaptic scaling^{1,2,17}. However, we have found that TNF- α is constitutively produced by glia⁸, suggesting that they, not neurons, might be the source of the TNF- α that is required for synaptic scaling in response to activity blockade. Glia are well placed to sense the overall activity of a network^{23,24} through glutamate spillover²⁵, direct synaptic contact²⁶ or from other activity-regulated signals, and then

provide a feedback signal to regulate network behaviour. To test directly the hypothesis that the TNF- α required for homeostatic synaptic strengthening is produced by glia, we plated either wild-type or *Tnf*^{-/-} neurons onto confluent beds of glia derived from either wild-type or *Tnf*^{-/-} hippocampi (Fig. 4a). Consistent with our earlier findings, *Tnf*^{-/-} neurons plated on *Tnf*^{-/-} glia did not show synaptic scaling in response to TTX treatment (Fig. 4c, d). However, *Tnf*^{-/-} neurons from the same culture preparation plated onto wild-type glia showed robust increases in mEPSC amplitude after the same TTX treatment (Fig. 4c, d). This strongly suggests that glial-derived TNF- α is sufficient to induce synaptic scaling in response to activity deprivation.

As neurons can also produce TNF- α , we then tested whether glial-derived TNF- α is required for the scaling up of excitatory synaptic

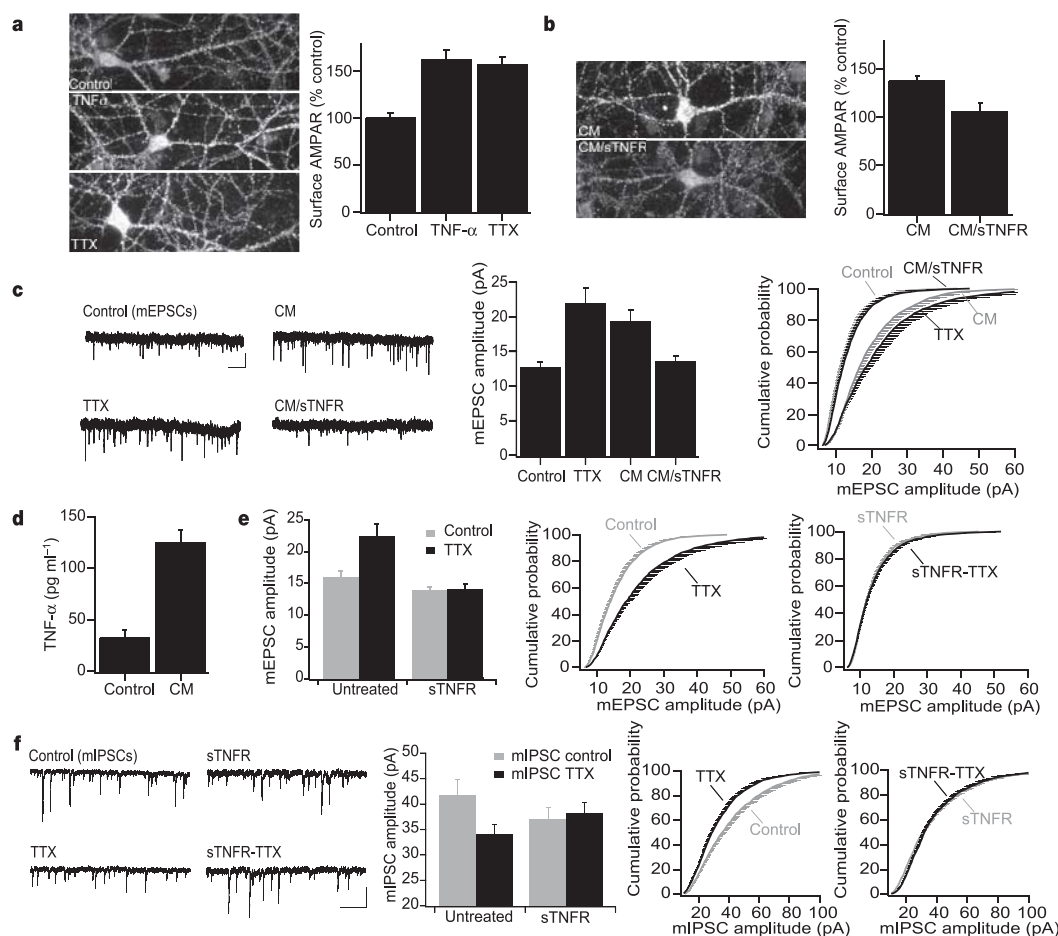


Figure 2 | TNF- α mediates synaptic scaling during activity blockade.

a, Representative micrographs and group data showing that acute TNF- α application or two days of TTX treatment increase AMPAR surface expression in cultured hippocampal neurons relative to untreated neurons (TNF- α , 162 $\pm$ 10% of control levels, $P < 0.0001$, $n = 72$; TTX, 157 $\pm$ 7%, $P < 0.005$, $n = 90$). **b**, Sample micrographs and group data showing that acute treatment with conditioned medium (CM) from cultures treated for two days with TTX increases AMPAR surface expression compared with untreated neurons (CM, 138 $\pm$ 5%, $P < 0.0001$, $n = 174$). However, conditioned medium incubated with sTNFR does not increase AMPAR surface expression (CM/sTNFR, 106 $\pm$ 8%, $P > 0.9$, $n = 75$). **c**, Left, examples of mEPSCs recorded from cultured neurons, either untreated (control), treated for two days with TTX, acutely treated with conditioned medium from the TTX-treated cultures (CM), or acutely treated with conditioned medium incubated with sTNFR (CM/sTNFR) (scale bars 20 pA, 1 s). Middle, histogram of average mEPSC amplitudes showing that TTX (22.0 $\pm$ 2.2 pA, $n = 11$ cells, $P < 0.0001$) and CM (19.4 $\pm$ 1.5 pA, $P < 0.002$, $n = 14$) but not CM/sTNFR (13.6 $\pm$ 0.7 pA, $P > 0.9$, $n = 11$) significantly increase mEPSC amplitude compared with control cells

(12.7 $\pm$ 0.7 pA, $n = 14$). Right, cumulative distributions of mEPSC amplitudes from cells shown in the histogram. **d**, TNF- α levels measured in media from control and TTX-treated cultures (control, 33 $\pm$ 7 pg ml⁻¹; CM, 125 $\pm$ 12 pg ml⁻¹; $n = 9$; $P < 0.0001$). **e**, Group data showing that addition of sTNFR during TTX treatment prevents the increase in mEPSC amplitude seen in TTX-treated sister cultures (control, 16.0 $\pm$ 1.0 pA, $n = 17$; TTX, 22.4 $\pm$ 2.0 pA, $n = 12$; sTNFR, 13.9 $\pm$ 0.6 pA, $n = 15$, $P < 0.05$; sTNFR-TTX, 14.1 $\pm$ 0.8 pA; $n = 16$, $P > 0.7$). Cumulative amplitude distributions of control (grey) and TTX-treated (black) cells, showing a rightward shift in mEPSC amplitudes attributable to TTX, and no shift in distribution of sTNFR-treated and sTNFR-TTX-treated cells. **f**, Sample recordings, average mIPSC amplitudes and cumulative amplitude distributions of mIPSCs from cells chronically treated with TTX, sTNFR or both, showing that inclusion of sTNFR also prevents the decrease in mIPSC amplitude observed after TTX treatment (TTX, 34.0 $\pm$ 2.0 pA; control, 41.8 $\pm$ 3.0 pA, $P < 0.05$, $n = 12$ –14; sTNFR-TTX, 38.2 $\pm$ 2.1 pA; sTNFR, 37.1 $\pm$ 2.3 pA, $P > 0.9$, $n = 12$ –14; scale bars 50 pA, 1 s). Error bars show s.e.m.

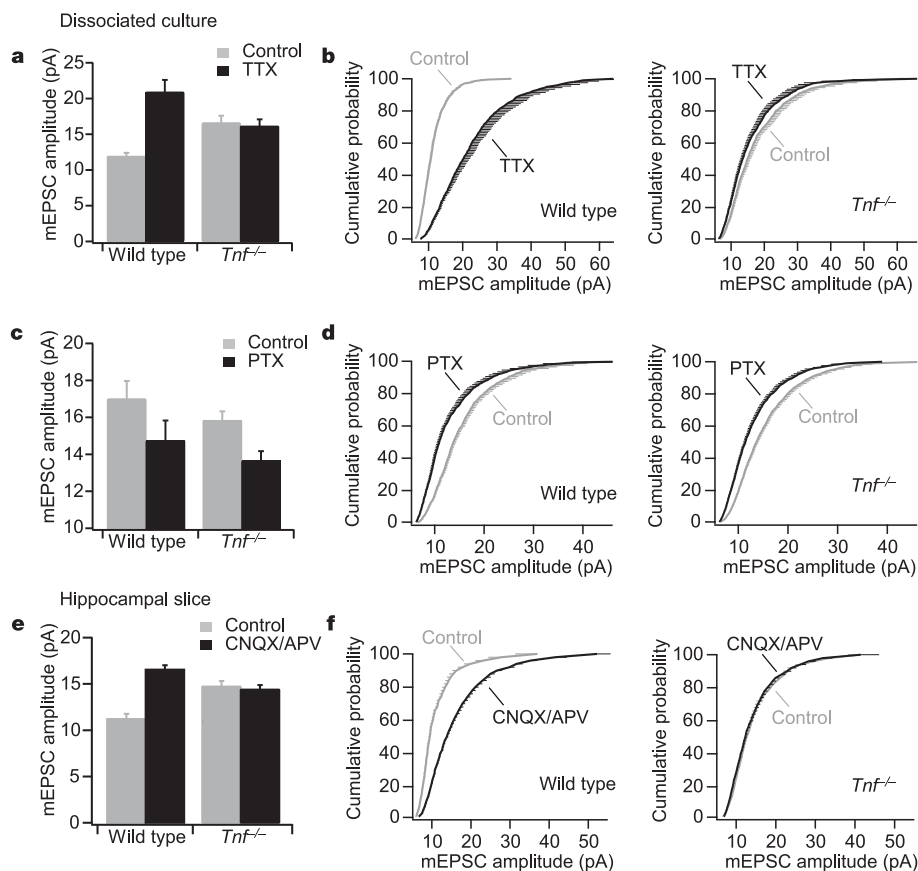


Figure 3 | *Tnf*^{-/-} neurons scale down but not up following activity manipulations. **a**, Mean mEPSC amplitude of control and TTX-treated neurons from wild-type mouse cultures (TTX, 20.9 ± 1.7 pA, $n = 12$; control, 11.9 ± 0.5 pA, $n = 12$; $P < 0.0001$) and *Tnf*^{-/-} cultures (TTX, 16.1 ± 0.9 pA, $n = 17$; control, 16.6 ± 1.0 pA, $n = 15$; $P > 0.7$). **b**, Cumulative mEPSC amplitude distributions of wild-type and *Tnf*^{-/-} mouse cultures for control (grey) or TTX-treated (black) cells. **c**, Mean mEPSC amplitude from control and PTX-treated neurons from wild-type mouse cultures (PTX, 14.7 ± 1.0 pA, $n = 13$; control, 17.0 ± 0.9 pA, $n = 12$; $P < 0.03$) and *Tnf*^{-/-} mouse cultures (PTX, 13.7 ± 0.5 pA, $n = 20$; control,

15.7 ± 0.6 pA, $n = 17$; $P < 0.002$). **d**, Cumulative mEPSC amplitude distributions from wild-type and *Tnf*^{-/-} cultures. **e**, Mean mEPSC amplitudes from hippocampal slice cultures of untreated neurons and neurons treated with the glutamate receptor antagonists CNQX and APV, in wild-type mice (CNQX/APV, 16.5 ± 0.5 pA, $n = 15$; control, 11.2 ± 0.6 pA, $n = 15$; $P < 0.0001$) and *Tnf*^{-/-} mice (CNQX/APV, 14.4 ± 0.5 pA, $n = 18$; control, 14.7 ± 0.6 pA, $n = 17$; $P > 0.6$). **f**, Cumulative mEPSC amplitude distributions from wild-type and *Tnf*^{-/-} mouse slice cultures. Error bars show s.e.m.

responses. As expected, wild-type neurons plated onto wild-type glia showed robust synaptic scaling (Fig. 4e, f). However, wild-type neurons from the same culture preparation plated onto *Tnf*^{-/-} glia no longer showed an increase in mEPSC amplitude in response to chronic activity blockade (Fig. 4e, f). Together, these data indicate that glial-derived TNF- α is both necessary and sufficient to induce the increase in mEPSC amplitude that results from chronic activity deprivation, and we therefore suggest that glial TNF- α is a critical signalling component for this form of synaptic scaling.

How might glia sense the overall activity of an adjacent neuronal network? A straightforward hypothesis is that glia are sensitive to glutamate, the level of which would correlate with local excitatory synaptic activity. As no neurons are present in glial-only cultures (which release significant levels of TNF- α ⁸), we reasoned that these cultures represent a zero-activity state. Using changes in the surface expression of AMPARs as a bioassay for TNF- α , we compared the effects of conditioned medium from glial cultures (GCM) to conditioned medium from sister glial cultures treated with glutamate (100 μ M) for the duration of the conditioning (48 h), a manipulation used to mimic exposure of glia to active neural networks. As we expected⁸, treatment of neuronal cultures with GCM increased cell-surface AMPAR levels (Fig. 4g). However, GCM from glutamate-treated glia was significantly less effective at increasing surface AMPAR levels (Fig. 4g). The amount of TNF- α in GCM was variable across preparations (340 ± 182 pg ml⁻¹, $n = 6$; data not shown),

as measured by an L929 cell death assay¹⁶, but again glutamate treatment consistently reduced TNF- α levels from individual preparations by about 50% (Fig. 4h). Thus, ambient glutamate might be one mechanism by which glia sense the level of neuronal activity and adjust TNF- α production.

Recent evidence suggests that glia act, at least in part, through the release of soluble factors to promote the formation of excitatory synapses during early development, and may also be important for their maintenance^{27,28}. Our data extend these observations by showing that glia are actively involved in mediating an important form of activity-dependent plasticity at established functioning synapses. Thus, through their release of different factors at different developmental stages (for example, thrombospondin release early during development²⁷ and TNF- α release at later times), glia may be actively involved in both the establishment of neural circuits as well as their subsequent activity-dependent regulation. Although we have not identified the subtype of glia that is responsible for TNF- α production, two likely sources include astrocytes and microglia (which typically make up <5% of the cells in the glial feeder layer in our cultures).

Combined with recent work on the neuronal role of MHC (major histocompatibility complex) class I molecules²⁹, our results showing a previously unrecognized neuronal function for constitutively released TNF- α suggest that, perhaps owing to the relative isolation of the nervous and immune systems, traditional immune signalling

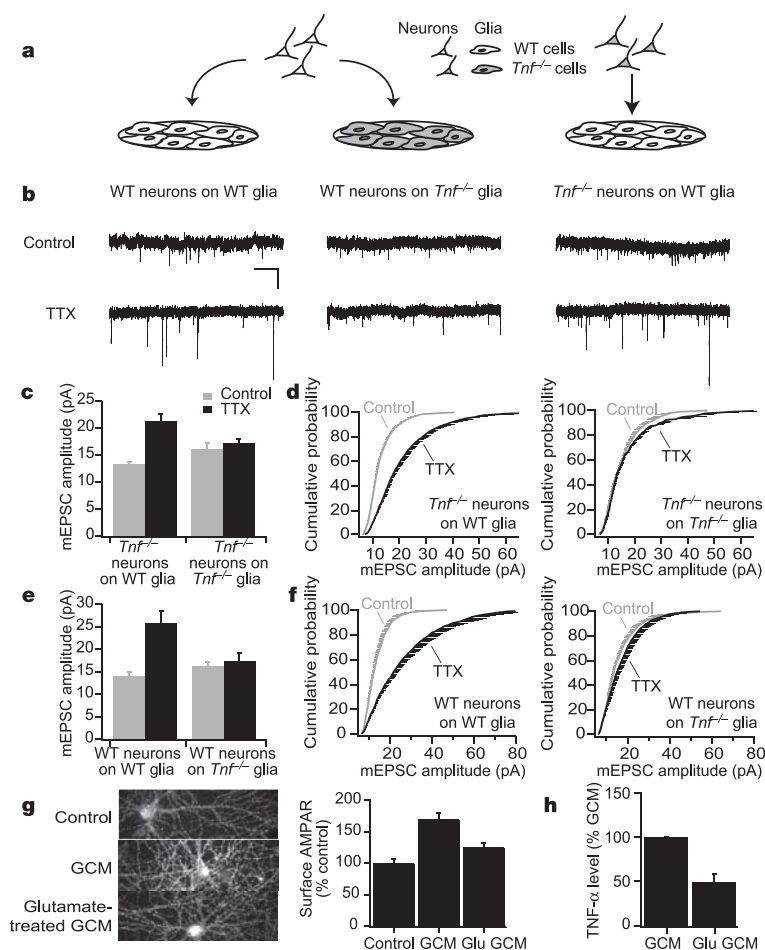


Figure 4 | Glia are the source of TNF- α required for synaptic scaling during activity blockade. **a**, Diagram illustrating how wild-type (WT) mouse neurons (white) were plated onto beds of either wild-type (white) or *Tnf*^{-/-} (grey) mouse glia. *Tnf*^{-/-} neurons (grey) were plated onto wild-type mouse glia. **b**, Representative recording of mEPSCs from control and TTX-treated cells, from neurons grown on different types of glia (scale bars 20 pA, 1 s). **c**, Mean amplitudes of mEPSCs recorded from *Tnf*^{-/-} neurons showing that TTX-induced increases in mEPSC amplitude occur when *Tnf*^{-/-} neurons are plated on wild-type glia (TTX, 21.2 $\pm$ 1.2 pA, *n* = 17; control, 13.2 $\pm$ 0.4 pA, *n* = 15; *P* < 0.0001) but not on *Tnf*^{-/-} glia (TTX, 17.2 $\pm$ 0.8 pA, *n* = 6; control, 16.0 $\pm$ 1.1 pA, *n* = 6, *P* > 0.4). **d**, Cumulative distribution of mEPSC amplitudes for *Tnf*^{-/-} neurons grown on wild-type or *Tnf*^{-/-} glia. **e**, Mean mEPSC amplitudes recorded

from wild-type neurons, showing that TTX-induced increases in mEPSC amplitude occur when wild-type neurons are plated on wild-type glia (TTX, 25.7 $\pm$ 2.8 pA, *n* = 10; control, 13.9 $\pm$ 1.1 pA, *n* = 9; *P* < 0.002) but not on *Tnf*^{-/-} glia (TTX, 17.3 $\pm$ 1.8 pA, *n* = 14; control, 16.1 $\pm$ 1.1 pA, *n* = 12; *P* > 0.6). **f**, Cumulative mEPSC amplitude distributions for control and TTX-treated wild-type neurons plated on wild-type or *Tnf*^{-/-} glia. **g**, Micrographs and group data showing that conditioned medium from glutamate-treated glia (glu GCM) is less effective than conditioned medium from untreated glia (GCM) at increasing cell-surface expression of AMPARs (glu GCM, 125 $\pm$ 7% of control cells; GCM, 170 $\pm$ 10% of control cells; *n* = 90; *P* < 0.01). **h**, Glutamate treatment also decreases TNF- α levels measured in GCM using an L929 cell death assay (49 $\pm$ 10% of untreated GCM, *n* = 4, *P* < 0.01). Error bars show s.e.m.

molecules have been adapted to serve different roles in nervous system function. Our data also suggest that changes in the levels of immune molecules in the brain during disease or in response to physical insults may lead to unexpected dysfunctions, because of both their direct neural actions and the disruption of their normal adaptive roles in the brain.

METHODS

See Supplementary Information for detailed methods.

Immunostaining and electrophysiology in cultures. Postnatal mixed hippocampal cultures were prepared as previously described⁹, and assayed at 13–15 days *in vitro*. For some experiments, mouse neurons were plated onto previously prepared confluent beds of 2–4-week-old mouse glia. FUDR (5-fluoro-2'-deoxyuridine) was added shortly after plating to prevent the proliferation of new glia. Surface AMPARs were visualized as described⁹. *n* values in the text represent the number of microscope fields examined. Statistical significance was determined by analysis of variance (ANOVA) followed by Dunnett's *t*-tests. Whole-cell patch-clamp recordings of mEPSCs were made essentially as described¹⁰. A positive control for synaptic scaling was performed on every culture batch. Cumulative distributions were generated using 200 consecutive

mEPSCs from each cell, averaged across all cells, and compared using a Kolmogorov–Smirnov two-sample test.

Slice electrophysiology. Acute transverse hippocampal slices (400 μ m) were prepared from 2–4-week-old Sprague-Dawley rats or mice as previously described⁹. LTP was induced with 1-s, 100-Hz trains, repeated three times at 20-s intervals. LTD was induced using 900 pulses at 3 Hz. Hippocampal slice cultures were prepared from postnatal day (P) 5–6 mice essentially as described³⁰, as modified by O. Schluter and W. Xu (personal communication). After 6–9 days *in vitro*, whole-cell recordings of mEPSCs from CA1 pyramidal cells were made, with the addition of TTX (500 nM) and sucrose (20 mM) to the external solution.

Received 25 October 2005; accepted 24 February 2006.

Published online 19 March 2006.

1. Turrigiano, G. G. & Nelson, S. B. Homeostatic plasticity in the developing nervous system. *Nature Rev. Neurosci.* 5, 97–107 (2004).
2. Burrone, J. & Murthy, V. N. Synaptic gain control and homeostasis. *Curr. Opin. Neurobiol.* 13, 560–567 (2003).
3. Abbott, L. F. & Nelson, S. B. Synaptic plasticity: taming the beast. *Nature Neurosci.* 3 (Suppl.), 1178–1183 (2000).
4. Malenka, R. C. & Bear, M. F. LTP and LTD: an embarrassment of riches. *Neuron* 44, 5–21 (2004).

5. O'Brien, R. J. *et al.* Activity-dependent modulation of synaptic AMPA receptor accumulation. *Neuron* **21**, 1067–1078 (1998).
6. Wierenga, C. J., Ibata, K. & Turrigiano, G. G. Postsynaptic expression of homeostatic plasticity at neocortical synapses. *J. Neurosci.* **25**, 2895–2905 (2005).
7. Kilman, V., van Rossum, M. C. & Turrigiano, G. G. Activity deprivation reduces miniature IPSC amplitude by decreasing the number of postsynaptic GABA_A receptors clustered at neocortical synapses. *J. Neurosci.* **22**, 1328–1337 (2002).
8. Beattie, E. C. *et al.* Control of synaptic strength by glial TNF α . *Science* **295**, 2282–2285 (2002).
9. Stellwagen, D., Beattie, E. C., Seo, J. Y. & Malenka, R. C. Differential regulation of AMPA receptor and GABA receptor trafficking by tumor necrosis factor- α . *J. Neurosci.* **25**, 3219–3228 (2005).
10. Carroll, R. C., Lissin, D. V., von Zastrow, M., Nicoll, R. A. & Malenka, R. C. Rapid redistribution of glutamate receptors contributes to long-term depression in hippocampal cultures. *Nature Neurosci.* **2**, 454–460 (1999).
11. Pasparakis, M., Alexopoulou, L., Episkopou, V. & Kollias, G. Immune and inflammatory responses in TNF α -deficient mice: a critical requirement for TNF α in the formation of primary B cell follicles, follicular dendritic cell networks and germinal centers, and in the maturation of the humoral immune response. *J. Exp. Med.* **184**, 1397–1411 (1996).
12. Peschon, J. J. *et al.* TNF receptor-deficient mice reveal divergent roles for p55 and p75 in several models of inflammation. *J. Immunol.* **160**, 943–952 (1998).
13. Ju, W. *et al.* Activity-dependent regulation of dendritic synthesis and trafficking of AMPA receptors. *Nature Neurosci.* **7**, 244–253 (2004).
14. Bezzi, P. *et al.* CXCR4-activated astrocyte glutamate release via TNF α : amplification by microglia triggers neurotoxicity. *Nature Neurosci.* **4**, 702–710 (2001).
15. Turrigiano, G. G., Leslie, K. R., Desai, N. S., Rutherford, L. C. & Nelson, S. B. Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature* **391**, 892–896 (1998).
16. Trost, L. C. & Lemasters, J. J. A cytotoxicity assay for tumor necrosis factor employing a multiwell fluorescence scanner. *Anal. Biochem.* **220**, 149–153 (1994).
17. Leslie, K. R., Nelson, S. B. & Turrigiano, G. G. Postsynaptic depolarization scales quantal amplitude in cortical pyramidal neurons. *J. Neurosci.* **21**, RC170 (2001).
18. Ganguly, K., Schinder, A. F., Wong, S. T. & Poo, M. GABA itself promotes the developmental switch of neuronal GABAergic responses from excitation to inhibition. *Cell* **105**, 521–532 (2001).
19. Rutherford, L. C., Nelson, S. B. & Turrigiano, G. G. BDNF has opposite effects on the quantal amplitude of pyramidal neuron and interneuron excitatory synapses. *Neuron* **21**, 521–530 (1998).
20. Shi, S. H. *et al.* Rapid spine delivery and redistribution of AMPA receptors after synaptic NMDA receptor activation. *Science* **284**, 1811–1816 (1999).
21. Tomita, S., Stein, V., Stocker, T. J., Nicoll, R. A. & Brecht, D. S. Bidirectional synaptic plasticity regulated by phosphorylation of stargazin-like TARPs. *Neuron* **45**, 269–277 (2005).
22. Hayashi, Y. *et al.* Driving AMPA receptors into synapses by LTP and CaMKII: requirement for GluR1 and PDZ domain interaction. *Science* **287**, 2262–2267 (2000).
23. Auld, D. S. & Robitaille, R. Glial cells and neurotransmission: an inclusive view of synaptic function. *Neuron* **40**, 389–400 (2003).
24. Aguado, F., Espinosa-Parrilla, J. F., Carmona, M. A. & Soriano, E. Neuronal activity regulates correlated network properties of spontaneous calcium transients in astrocytes *in situ*. *J. Neurosci.* **22**, 9430–9444 (2002).
25. Diamond, J. S. Deriving the glutamate clearance time course from transporter currents in CA1 hippocampal astrocytes: transmitter uptake gets faster during development. *J. Neurosci.* **25**, 2906–2916 (2005).
26. Lin, S. C. & Bergles, D. E. Synaptic signaling between neurons and glia. *Glia* **47**, 290–298 (2004).
27. Christopherson, K. S. *et al.* Thrombospondins are astrocyte-secreted proteins that promote CNS synaptogenesis. *Cell* **120**, 421–433 (2005).
28. Ullian, E. M., Sapperstein, S. K., Christopherson, K. S. & Barres, B. A. Control of synapse number by glia. *Science* **291**, 657–661 (2001).
29. Huh, G. S. *et al.* Functional requirement for class I MHC in CNS development and plasticity. *Science* **290**, 2155–2159 (2000).
30. Stoppini, L., Buchs, P. A. & Muller, D. A simple method for organotypic cultures of nervous tissue. *J. Neurosci. Methods* **37**, 173–182 (1991).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Wu for expert technical assistance, J. Tsui and B. Stevens for advice on tissue culture techniques, W. Xu, P. Luu and W. Morishita for discussions, and R. Huganir and B. Barres for reagents. This work was funded by NIH grants to R.C.M. and D.S.

Author Contributions D.S. performed the experiments and data analyses. D.S. and R.C.M. designed the experiments, discussed the interpretation of results and co-wrote the paper.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.C.M. (malenka@stanford.edu).

LETTERS

Unique features of action potential initiation in cortical neurons

Björn Naundorf^{1,2,3}, Fred Wolf^{1,2,3} & Maxim Volgushev^{4,5}

Neurons process and encode information by generating sequences of action potentials^{1,2}. For all spiking neurons, the encoding of single-neuron computations into sequences of spikes is biophysically determined by the cell's action-potential-generating mechanism. It has recently been discovered that apparently minor modifications of this mechanism can qualitatively change the nature of neuronal encoding^{3,4}. Here we quantitatively analyse the dynamics of action potential initiation in cortical neurons *in vivo*, *in vitro* and in computational models. Unexpectedly, key features of the initiation dynamics of cortical neuron action potentials—their rapid initiation and variable onset potential—are outside the range of behaviours described by the classical Hodgkin–Huxley theory. We propose a new model based on the cooperative activation of sodium channels that reproduces the observed dynamics of action potential initiation. This new model predicts that Hodgkin–Huxley-type dynamics of action potential initiation can be induced by artificially decreasing the effective density of sodium channels. *In vitro* experiments confirm this prediction, supporting the hypothesis that cooperative sodium channel activation underlies the dynamics of action potential initiation in cortical neurons.

We analysed action potentials elicited in cortical neurons *in vivo* and *in vitro*, either spontaneously or in response to various stimuli. In all cells examined and for all conditions tested, the dynamics of action potential initiation was characterized by a very abrupt onset and a steep upstroke in membrane potential. In membrane potential traces, the abrupt onset of action potentials is apparent as a sharp kink (Fig. 1a, c). This phenomenon stands out even more clearly in phase plots that graph the rate of change of the membrane potential dV/dt against the instantaneous membrane potential $V(t)$, and is manifested as an almost vertical take-off in dV/dt versus V trajectories at action potential onset (Fig. 1b, d). In phase plots, an action potential is represented by a loop. At the start of the loop, the velocity increases rapidly from less than 5 mV ms^{-1} to more than 20 mV ms^{-1} . This several-fold increase in velocity occurs within a range of less than 1 mV and takes less than 0.2 ms. This onset behaviour is not a peculiarity of neurons *in vivo*. Neurons *in vitro* show similarly fast dynamics of action potential initiation (Fig. 1c, d). These dynamic features of recorded action potential onsets distinguish them from the behaviour of previously proposed computational models. Figure 1e, f shows a simulated action potential using a recently developed conductance-based model of a cortical neuron⁵. In this model, a velocity of 20 mV ms^{-1} is only reached over a range of 7–8 mV after about 1 ms. Thus, the real onset of cortical action potentials is approximately ten times faster than predicted by the model.

The rapid onset of action potentials is a very robust phenomenon, apparent during spontaneous and evoked activity *in vivo*. Moreover,

it is independent of the temporal structure of synaptic inputs (Fig. 2) and of the electrophysiological cell class (Fig. 4). Fig. 2 shows phase plots and membrane potential traces from a simple cell (Fig. 2a, b) and a complex cell (Fig. 2c, d) recorded *in vivo* in cat visual cortex. In the phase plots, subthreshold membrane potential fluctuations are represented by a grey cloud. In both the simple and the complex cell, action potentials rise almost vertically out of this cloud. Detecting the point at which the rate of change reaches a value of 10 mV ms^{-1} thus allows reliable identification of the time of action potential initiation

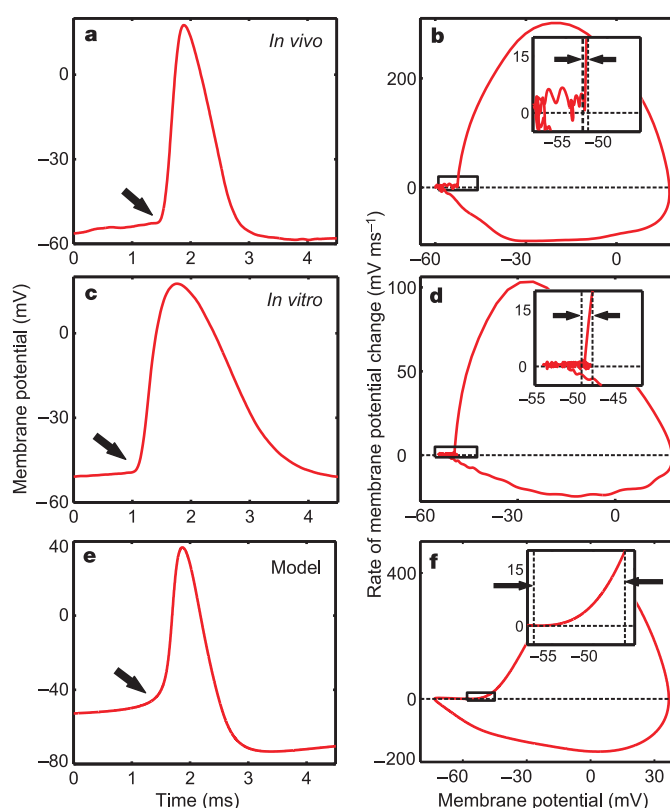


Figure 1 | Dynamics of action potential initiation in neocortical neurons and in a Hodgkin–Huxley-type model of a neocortical neuron. **a**, Action potential in a cat visual cortex neuron *in vivo*. The arrow shows the characteristic kink at action potential onset. **b**, Phase plot (dV/dt versus V) of the action potential from **a**. Inset shows the initial phase of the action potential. **c**, **d**, Action potential from a cat visual cortex slice *in vitro* at 20°C . **e**, **f**, Action potential from a Hodgkin–Huxley-type model of a neocortical neuron⁵.

¹Max Planck Institute for Dynamics and Self-Organization, ²Department of Physics and ³Bernstein Center for Computational Neuroscience, University of Göttingen, Bunsenstr. 10, D-37073 Göttingen, Germany. ⁴Department of Neurophysiology, Ruhr-University Bochum, D-44780 Bochum, Germany. ⁵Institute of Higher Nervous Activity and Neurophysiology Russian Academy of Sciences, Moscow 117485, Russia.

and the onset potential. The phase plots also show a second salient feature of cortical action potentials: the onset potentials vary considerably, over ranges of up to 10 mV (Fig. 2a–d) (Special care was taken to exclude any non-stationarities, see Supplementary Information). This distinct variability in onset potentials has previously been observed in cat visual cortex^{6–8} and rat hippocampus⁹. This second feature of cortical action potentials is also missing in Hodgkin–Huxley-type models. Figure 2e, f depicts the behaviour of such a model driven by fluctuating synaptic inputs⁵. The variability in onset potentials in this model is restricted to a range of less than 2 mV, which is much smaller than observed *in vivo*.

Two features thus render cortical action potentials distinctly different from simulated action potentials using Hodgkin–Huxley-type models. First, the initial action potential phase is approximately ten times faster in recorded neurons compared to conductance-based models. Second, the onset potential variability is approximately five times larger in the recorded cells. We tried, using various modifications of the models, to achieve a better match between recorded and simulated action potentials (including and/or modifying adaptation currents¹⁰, channel stochasticity¹¹, state-dependent inactivation¹², sodium channel activation curves and peak conductances; see Supplementary Information). None of the modified models reproduced the two salient features of the recorded action potentials.

In fact, a straightforward analysis reveals that rapid action potential onset and large variability in onset potentials are strongly antagonistic in Hodgkin–Huxley-type models. In such models, the initial phase of an action potential is determined by the activation of

voltage-dependent sodium channels. Their dynamics is described by the activation curve and kinetics of an associated gating variable. In the Hodgkin–Huxley formulation it can be shown that the rate of membrane depolarization is limited by $g_{\text{Na}}h_0m_{\infty}^3(V)(V_{\text{Na}} - V)/C + I_0/C$, where g_{Na} denotes peak sodium conductance, h_0 is the fraction of sodium channels available for activation, $m_{\infty}^3(V)$ is their activation curve, V_{Na} is the sodium reversal potential, C the membrane capacitance, and I_0 is the current carried by other channels. This upper bound on the rate of membrane potential change links the action potential onset dynamics directly to the width of the activation curve and peak sodium conductance. Figure 3 illustrates this relationship. An experimentally obtained activation curve from patches of cortical neurons^{13,14} results in a shallow action potential onset (Fig. 3a, b). Increasing the steepness of the activation curve leads to sharper action potential onsets, but even with a fivefold increase, the simulated action potentials do not rise as fast as those recorded. Changing the effective peak sodium conductance—mimicking inactivation^{6,9}—leaves the steepness of action potential onset unaffected but shifts the onset potentials (Fig. 3c, d). At the same time, increasing the steepness of the activation curve considerably decreases onset potential variability (Fig. 3c–f). Quantitatively, the variability of onset potentials is restricted by $\Delta V \approx k \log G$, where k is the width of the activation curve and G is the ratio of maximum to minimum peak conductance (see Supplementary Information). To mimic the measured combinations of onset rapidness and variability, unphysiologically large values of G (about 20,000) would be required.

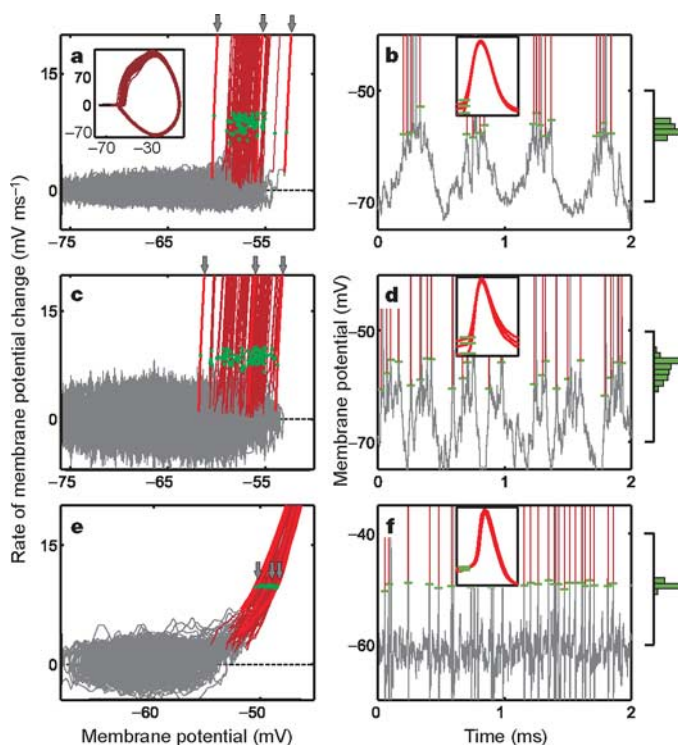


Figure 2 | Different action potential initiation in visual cortex neurons recorded *in vivo* and in a Hodgkin–Huxley-type model subject to fluctuating synaptic inputs. **a**, Phase plot of a simple cell response to a moving grating of optimal orientation. Subthreshold fluctuations are shown in grey, action potentials in red, and green dots indicate action potential onsets. The inset shows the complete trace. Arrows indicate three sample action potentials. **b**, Part of the recording from **a**, using the same colour code, with action potentials truncated in amplitude. Green bars show action potential onset potentials. Inset shows the action potentials marked with arrows in **a**. The histogram to the right shows the distribution of action potential onset potentials. **c**, **d**, Response of a complex cell. **e**, **f**, Response of a Hodgkin–Huxley-type model⁵ subject to fluctuating synaptic input.

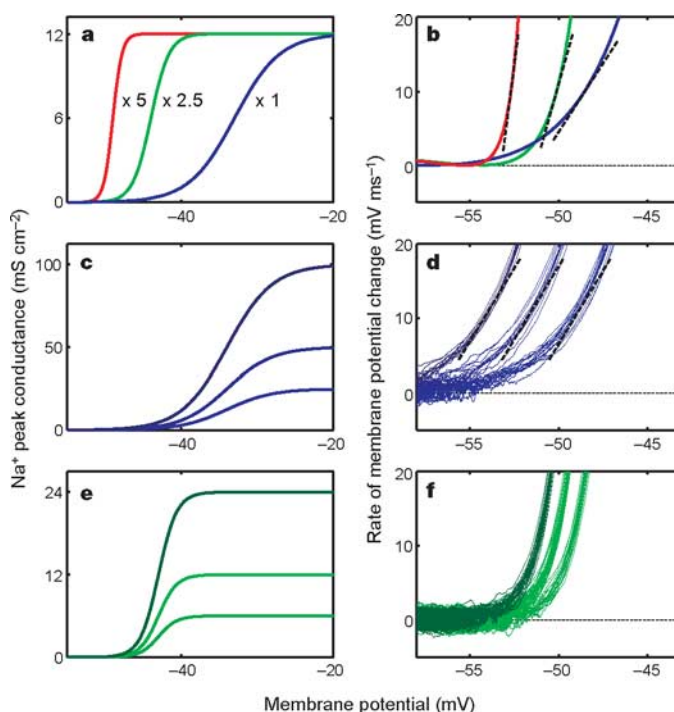


Figure 3 | Effect of the shape of the sodium channel activation curve and effective peak conductance on action potential initiation in a Hodgkin–Huxley-type model. Activation curves used in the model (**a**, **c**, **e**) and initial phases of resulting action potentials (**b**, **d**, **f**) are shown using matched colours. Blue lines show standard activation curves¹⁴. **a**, **b**, Increasing the steepness of the activation curve (**a**) leads to a more rapid action potential upstroke (**b**). Dashed lines indicate tangents at 10 mV ms^{-1} . **c**–**f**, Initiation of action potentials with shallow (**c**, **d**) and steep (**e**, **f**) activation curves and different sodium peak conductances in a model driven by fluctuating synaptic inputs (several action potentials superimposed). Changing peak conductance shifts the action potential onset potential but does not affect its onset rapidness. Steeper activation curves lead to smaller action potential onset spans (**d**, 6 mV; **f**, 2.5 mV).

To quantitatively compare the action potential onset dynamics in our recordings with action potential dynamics in Hodgkin–Huxley-type models, we plotted the action potential onset span (difference between maximum and minimum onset potential in a recording) against the rapidness of action potential onset (the slope of the phase plot at $dV/dt = 10 \text{ mV ms}^{-1}$) for real and simulated recordings (Fig. 4). In the simulations, we used two different models^{5,10} driven by fluctuating synaptic currents. In both models we systematically changed the peak sodium conductance and the activation curve over the entire range in which action potentials were generated. The locations of data points from the model simulations reflect the antagonism between onset span and rapidness. Simulated action potentials either showed a large onset rapidness or a large onset span, but never both. The points representing simulated action potentials are clearly separated from the points representing *in vivo* action potentials, which show rapid onset dynamics and large variability in onset potentials, irrespective of the electrophysiological cell type. Action potentials recorded *in vitro* had similarly fast onset dynamics (Fig. 4).

The above arguments and our extensive simulations indicate that the dynamics of action potential initiation in cortical neurons deviates qualitatively from the classical picture described by the Hodgkin–Huxley framework. What could be the biophysical mechanism that enables cortical action potentials to initiate much faster and at the same time with a much larger onset potential variability than predicted by the Hodgkin–Huxley theory? According to this theory, there is a one-to-one relationship between the single-channel activation curve and the action potential onset dynamics, owing to the assumption that the opening of individual sodium channels is statistically independent. This assumption, however, might be violated in the highly organized molecular machinery of a living

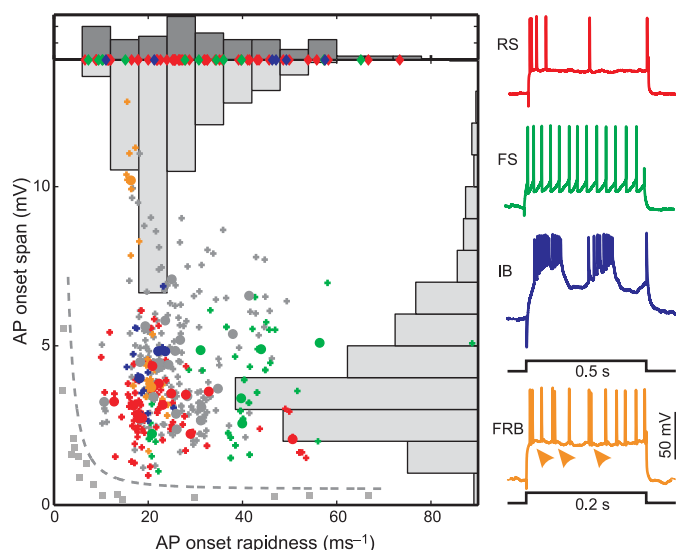


Figure 4 | Action potential onset span and rapidness in cortical neurons and Hodgkin–Huxley-type models. Points show data from cat visual cortex neurons *in vivo*, classified electrophysiologically as regular spiking (RS), fast spiking (FS), intrinsically bursting (IB) or fast rhythmic bursting (FRB). Colours match the neuronal responses to current steps shown on the right. Grey points indicate electrophysiologically unidentified cells. Circles show the mean of several measurements for each cell (individual measurements indicated with crosses). Diamonds at the top of the panel show *in vitro* data (cat visual cortex in blue, rat visual cortex in red, mouse hippocampus in green). Grey squares show simulation results from two Hodgkin–Huxley-type models with varied sodium channel activation curves and peak conductances, driven by fluctuating synaptic inputs. Dashed grey line separates the model from experimentally derived data. Histograms show the marginal distributions of the *in vivo* (light grey bars) and *in vitro* (dark grey bars) data.

cell. Indeed, the rapid onset of action potentials suggests that many sodium channels open virtually simultaneously, that is, in a potentially cooperative fashion.

To assess whether cooperative activation of voltage-gated sodium channels can account for the two characteristic features of cortical action potential initiation, we constructed a model of a population of coupled sodium channels. In this model, the gating of individual sodium channels follows a scheme introduced by Aldrich, Corey, and Stevens¹⁵. It incorporates state-dependent inactivation from the open state and voltage-dependent inactivation from closed states^{15,16}. The key feature of our model is a coupling between neighbouring channels: the opening of a channel shifts the activation curve of each channel to which it is coupled towards more hyperpolarized values, thus increasing its probability of opening.

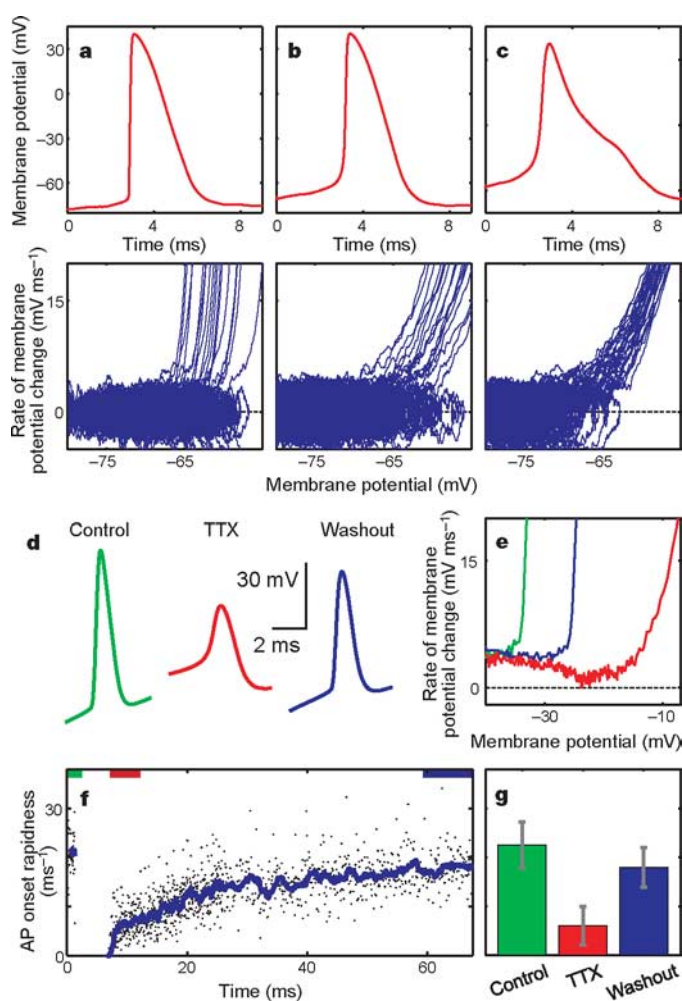


Figure 5 | Cooperative activation of voltage-gated sodium channels can account for the dynamics of action potential initiation in cortical neurons. **a**, Waveform (top) and phase plot (bottom) of action potentials elicited by fluctuating inputs in a conductance-based model that incorporates cooperative activation of sodium channels and closed-state inactivation. Both the action potential onset potential variability and onset rapidness are comparable to the *in vivo* recordings in Fig. 2. **b**, **c**, Same model, but without inter-channel coupling (**b**), and with Hodgkin–Huxley-like channel activation (**c**). **d–g**, Reducing the effective density of available sodium channels through TTX application reversibly reduces action potential amplitude and onset rapidness in cortical neurons *in vitro*. Shown are action potential waveforms (**d**) and phase plots of their initial parts (**e**). **f**, Time course of action potential onset rapidness in a cortical neuron before, during and after TTX application. **g**, Reversible reduction in onset rapidness of action potentials by TTX in six neurons. Error bars indicate s.e.m.

This model is able to reproduce the key features of cortical action potential initiation (Fig. 5a–c). With strongly cooperative activation, voltage-dependent inactivation from closed states, and slow de-inactivation (recovery from inactivation) of sodium channels, the simulated action potentials show both a large onset rapidness and large variability in onset potentials (Fig. 5a). Turning off the inter-channel coupling made the onset dynamics much shallower, while leaving onset variability unaffected (Fig. 5b). Hodgkin–Huxley-type dynamics of action potential onset was recovered when inactivation and de-inactivation were set to be fast and voltage-independent (Fig. 5c).

Assuming that channel interactions are distance-dependent in neuronal membranes, our model predicts that reducing the effective density of channels should weaken cooperativity, reduce the action potential onset rapidness and eventually lead to Hodgkin–Huxley-type onset dynamics. We tested this prediction *in vitro*, recording action potentials while reducing the density of available sodium channels by the application of tetrodotoxin (TTX). As expected, TTX application led to a decrease in action potential amplitude (Fig. 5d). More importantly, it also led to a substantial reduction in the onset rapidness of action potentials (Fig. 5d, e) in all tested cortical neurons (Fig. 5g), as predicted by our model. Moreover, gradual recovery of the number of available sodium channels during washout of TTX led to a gradual increase in action potential onset rapidness (Fig. 5f). These results cannot be explained by Hodgkin–Huxley-type models, in which reduction in the sodium channel density modifies only the amplitude of action potentials and their onset potential, but not their onset rapidness (see Fig. 3). Thus, in our opinion the fact that the dynamics of action potential initiation deviates qualitatively from voltage-dependent single channel activation points towards the cooperative activation of voltage-gated sodium channels.

Although our results are unexpected from a biophysical perspective, the combination of rapid dynamics and variable onset potentials of action potentials is beneficial for the coding of fast signals^{3,4} (see also Supplementary Information). With Hodgkin–Huxley-type dynamics of action potential initiation, the encoding of signals that vary on a timescale of less than 10 ms requires unphysiologically high mean firing rates that are likely to be energetically prohibitive^{17,18}. With action potential onset dynamics as described here for cortical neurons, much lower mean firing rates can support the encoding of such rapidly varying signals.

METHODS

In vivo and in vitro experiments. *In vivo* intracellular recordings were made using sharp electrodes in adult cats (3.0–4.5 kg). Data from 47 cells were used for the analysis. In each cell, we recorded responses to the presentation of moving gratings of different orientations (duration 5–7 s), and periods of spontaneous activity (10–120 s). Cells were classified functionally as either simple or complex using the spike response modulation index, and electrophysiologically by their responses to depolarizing current steps. *In vitro*, whole-cell recordings were made with patch electrodes in slices of rat or cat visual cortex and rat or mouse hippocampus. Data from 17 rat, 3 mouse and 2 cat neurons were analysed.

Computational models. We used two conductance-based Hodgkin–Huxley-type models^{5,10}, constructed to match the subthreshold membrane potential dynamics and firing statistics of cortical neurons. We introduced modifications to the models and varied model parameters over wide ranges in an attempt to reproduce the experimentally observed dynamics of action potential initiation. Action potential initiation by cooperative sodium channel activation was modelled using an effective mean field dynamics for a population of interacting sodium channels.

Details of experimental procedures and data analysis, and definitions of the models, their modifications and parameter ranges are provided in the Supplementary Information.

Received 22 October 2005; accepted 27 January 2006.

- Bernstein, J. Ueber den zeitlichen Verlauf der negativen Schwankung des Nervenstroms. *Pflügers Arch.* **1**, 173–207 (1868).
- Huxley, A. F. *Nobel Lectures, Physiology or Medicine 1963–1970* (Elsevier, Amsterdam, 1972).
- Fourcaud-Trocme, N., Hansel, D., van Vreeswijk, C. & Brunel, N. How spike generation mechanisms determine the neuronal response to fluctuating inputs. *J. Neurosci.* **23**, 11628–11640 (2003).
- Naundorf, B., Geisel, T. & Wolf, F. Action potential onset dynamics and the response speed of neuronal populations. *J. Comput. Neurosci.* **18**, 297–309 (2005).
- Destexhe, A. & Paré, D. Impact of network activity on the integrative properties of neocortical pyramidal neurons *in vivo*. *J. Neurophysiol.* **81**, 1531–1547 (1999).
- Azouz, R. & Gray, C. M. Dynamic spike threshold reveals a mechanism for synaptic coincidence detection in cortical neurons *in vivo*. *Proc. Natl Acad. Sci. USA* **97**, 8110–8115 (2000).
- Volgushev, M., Pernberg, J. & Eysel, U. T. A novel mechanism of response selectivity of neurons in cat visual cortex. *J. Physiol. (Lond.)* **540**, 307–320 (2002).
- Azouz, R. & Gray, C. M. Adaptive coincidence detection and dynamic gain control in visual cortical neurons *in vivo*. *Neuron* **37**, 513–532 (2003).
- Henze, D. A. & Buzsáki, G. Action potential threshold of hippocampal pyramidal cells *in vivo* is increased by recent spiking activity. *Neuroscience* **105**, 121–130 (2001).
- Wang, X. J., Liu, Y., Sanchez-Vives, M. V. & McCormick, D. A. Adaptation and temporal decorrelation by single neurons in the primary visual cortex. *J. Neurophysiol.* **89**, 3279–3293 (2003).
- Schneidman, E., Freedman, B. & Segev, I. Ion channel stochasticity may be critical in determining the reliability and precision of spike timing. *Neural Comput.* **10**, 1679–1703 (1998).
- Patlak, J. Molecular kinetics of voltage-dependent Na⁺ channels. *Physiol. Rev.* **71**, 1047–1080 (1991).
- Huguenard, J. R., Hamill, O. P. & Prince, D. A. Developmental changes in Na⁺ conductances in rat neocortical neurons: appearance of a slowly inactivating component. *J. Neurophysiol.* **59**, 778–795 (1988).
- Colbert, C. M. & Pan, E. Ion channel properties underlying axonal action potential initiation in pyramidal neurons. *Nature Neurosci.* **5**, 533–538 (2002).
- Aldrich, R. W., Corey, D. P. & Stevens, C. F. A reinterpretation of mammalian sodium channel gating based on single channel recording. *Nature* **306**, 436–441 (1983).
- Goldman, L. Stationarity of sodium channel gating kinetics in excised patches from neuroblastoma N1E 115. *Biophys. J.* **69**, 2364–2368 (1995).
- Attwell, D. & Laughlin, S. B. An energy budget for signaling in the grey matter of the brain. *J. Cereb. Blood Flow Metab.* **21**, 1133–1145 (2001).
- Lennie, P. The cost of cortical computation. *Curr. Biol.* **13**, 493–497 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We would like to thank A. Borst, M. Brecht, M. Chistiakova, T. Geisel, T. Kottos, T. Moser, E. Neher, W. Stühmer, I. Timofeev and C. van Vreeswijk for discussions, A. Borst, M. Brecht and E. Neher for comments on earlier versions of the manuscript, and A. Malyshev for help in some of the experiments. This study was supported by grants from the Deutsche Forschungsgemeinschaft to M.V., by grants from the Human Frontier Science Program and the Bundesministerium für Bildung und Forschung to F.W., and by the Max-Planck Society.

Author Contributions B.N., F.W. and M.V. contributed equally to this work. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to F.W. (fred@chaos.gwdg.de).

LETTERS

Enhanced bacterial clearance and sepsis resistance in caspase-12-deficient mice

Maya Saleh^{1†}, John C. Mathison², Melissa K. Wolinski¹, Steve J. Bensinger¹, Patrick Fitzgerald¹, Nathalie Droin¹, Richard J. Ulevitch², Douglas R. Green^{1*†} & Donald W. Nicholson^{*3}

Caspases function in both apoptosis and inflammatory cytokine processing and thereby have a role in resistance to sepsis¹. Here we describe a novel role for a caspase in dampening responses to bacterial infection. We show that in mice, gene-targeted deletion of caspase-12 renders animals resistant to peritonitis and septic shock. The resulting survival advantage was conferred by the ability of the caspase-12-deficient mice to clear bacterial infection more efficiently than wild-type littermates. Caspase-12 dampened the production of the pro-inflammatory cytokines interleukin (IL)-1 β , IL-18 (interferon (IFN)- γ inducing factor) and IFN- γ , but not tumour-necrosis factor- α and IL-6, in response to various bacterial components that stimulate Toll-like receptor and NOD pathways. The IFN- γ pathway was crucial in mediating survival of septic caspase-12-deficient mice, because administration of neutralizing antibodies to IFN- γ receptors ablated the survival advantage that otherwise occurred in these animals. Mechanistically, caspase-12 associated with caspase-1 and inhibited its activity. Notably, the protease function of caspase-12 was not necessary for this effect, as the catalytically inactive caspase-12 mutant Cys299Ala also inhibited caspase-1 and IL-1 β production to the same extent as wild-type caspase-12. In this regard, caspase-12 seems to be the cFLIP counterpart for regulating the inflammatory branch of the caspase cascade. In mice, caspase-12 deficiency confers resistance to sepsis and its presence exerts a dominant-negative suppressive effect on caspase-1, resulting in enhanced vulnerability to bacterial infection and septic mortality.

To address the function of caspase-12 in apoptosis and immunity, caspase-12-null mutant C57Bl/6 mice were generated in which a neo/LacZ cassette replaced a segment of exon 2 of the caspase-12 gene (*Casp12*), shifting the downstream sequence out of the reading frame (see Methods and Supplementary Fig. 1a, b). Caspase-12 transcripts and protein were not detected in mutant mice by either polymerase chain reaction (not shown) or western blotting (Supplementary Fig. 1c). Deletion of *Casp12* had no effect on the expression of neighbouring caspase genes in the inflammatory gene cluster—we observed comparable levels of lipopolysaccharide (LPS)-inducible caspase-11 transcripts and protein in the peripheral blood of wild-type and caspase-12-deficient mice (Supplementary Fig. 1d, e), and the expression of caspase-1 was similarly not altered by the deletion of *Casp12* (Supplementary Fig. 1f). As found for mice deficient in other inflammatory caspases (*Casp1*^{-/-} (refs 2, 3) and *Casp11*^{-/-} mice⁴), caspase-12-deficient (*Casp12*^{-/-}) mice were indistinguishable in phenotype compared to age- and gender-matched wild-type littermates (physical examination, necropsy, histological evaluation, blood counts and chemistry, and behavioural and fertility tests; data not shown).

The association of human caspase-12 (the caspase-12 long variant, CASP12L) with susceptibility to sepsis in individuals of African descent⁵ led us to examine the response of the *Casp12*^{+/+} and *Casp12*^{-/-} mice in a murine model of polymicrobial sepsis. To this end, we applied the colon ascendens stent peritonitis (CASP)⁶ model that closely mimics human sepsis. Whereas most of the wild-type mice succumbed to sepsis in the first 48 h after the peritonitis surgery, 60% of the *Casp12*^{-/-} mice were resistant to sepsis and survived (log rank test, $P = 0.00065$) (Fig. 1a), indicating that the presence of caspase-12 predisposed the mice to sepsis and that its absence conferred protection. Surprisingly, the *Casp12*^{-/-} mice were sensitive to LPS overdose: 90% of the knockout mice died by 72 h after LPS injection (not shown). This finding contrasts with the resistance to LPS overdose that was previously reported for *Casp1*^{-/-} (refs 2, 3) and *Casp11*^{-/-} mice⁴, indicating that caspase-12 function *in vivo* is not redundant with that of these two related caspases.

To dissect the mechanism responsible for the survival advantage of the caspase-12-deficient mice in this model, we measured blood bacterial content 12 h after CASP surgery. *Casp12*^{-/-} mice showed a markedly lower number of bacterial colony-forming units (c.f.u.) per ml blood compared to wild-type littermates (Fig. 1b, c), suggesting that more efficient bacterial clearance occurs in the absence of caspase-12. We evaluated the number and type of commensal bacteria from both mouse genotypes, and did not find any significant difference between the microflora of the *Casp12*^{-/-} compared to the wild-type mice (not shown). Therefore, the effect that we observed is most probably due to a difference in bacterial clearance. To determine whether this effect was restricted to abdominal infections or was a general characteristic of the caspase-12-deficient mice, a systemic route of infection was examined. *Casp12*^{+/+} and *Casp12*^{-/-} mice were challenged intravenously with a defined dose of the Gram-positive bacterium *Listeria monocytogenes*. Pathogen load was assessed by measuring c.f.u. in the spleen, liver and peripheral blood at varying time points after infection. Similar to what we observed in the CASP model, *Casp12*^{-/-} mice cleared *L. monocytogenes* more efficiently than wild-type mice (Fig. 1d, e, and data not shown). Taken together, these results suggest that caspase-12 has a critical role in regulating the clearance of bacterial pathogens such that in the absence of caspase-12, both systemic and abdominal infections are better contained and resolved.

We next sought to probe the signal transduction pathways and cytokine responses that might be affected by murine caspase-12. We chose to examine this in *ex vivo* splenocytes because we observed that the *lacZ* gene that was introduced into the caspase-12 targeted locus was highly induced by bacterial infection in the spleen (Fig. 2a). Splenocytes from *Casp12*^{+/+} and *Casp12*^{-/-} mice were isolated and

¹Department of Cellular Immunology, La Jolla Institute for Allergy and Immunology, San Diego, California 92121, USA. ²Department of Immunology, The Scripps Research Institute, San Diego, California 92037, USA. ³Merck Research Laboratories, Rahway, New Jersey 07065-0900, USA. [†]Present addresses: Department of Medicine, McGill University, Montreal H3A 1A1, Canada (M.S.); Department of Immunology, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA (D.R.G.).

*These authors contributed equally to this work.

analysed for their cytokine response to various Toll-like receptor (TLRs) ligands and the NOD2 ligand muramyl dipeptide (MDP). Notably, in response to some of the TLR/NOD ligands tested the caspase-12-deficient splenocytes produced more of the pro-inflammatory cytokines IL-1 β , IL-18 (also known as IGIF) and IFN- γ compared to wild-type cells (Fig. 2b), but not tumour-necrosis factor (TNF)- α or IL-6 (Supplementary Fig. 2). As a control, we examined the relative expression of pro-IL-1 β transcripts from *Casp12*^{+/+} and *Casp12*^{-/-} macrophages in response to TLR ligands to determine whether the negative effect of caspase-12 on IL-1 β production is due to inhibition of pro-IL-1 β transcription. Our results indicate that the levels of pro-IL-1 β message were not significantly different between the two mouse genotypes (Fig. 2c).

The lack of effect of caspase-12 status on TNF and IL-6 production was in contrast to what was observed in humans with CASP12L, the expression of which dampened the levels of TNF as well as that of granulocyte-macrophage colony stimulating factor (GM-CSF) and

IL-8 (ref. 5). It is therefore possible that murine caspase-12 and the human CASP12L variant may use non-overlapping as well as common signal transduction pathways in sepsis to regulate the inflammatory response. For example, whereas human CASP12L blocked NF- κ B signalling in response to TNF or LPS, rodent caspase-12 had no effect on NF- κ B regulation (our previous results⁵ and Supplementary Fig. 3).

In both humans and mice, IFN- γ is an established survival factor in sepsis⁶⁻⁸. Because the *Casp12*^{-/-} mice produced noticeably more IFN- γ in response to bacteria than wild-type mice, we tested whether this hyper-production of IFN- γ might account for the protection of the knockout mice from septic shock. Neutralizing antibodies to murine IFN- γ receptors were injected intraperitoneally in the *Casp12*^{-/-} mice 1 h before initiation of sepsis by CASP and survival was assessed as described above. Disruption of IFN- γ signalling ablated the survival advantage of caspase-12-deficient mice after CASP (Fig. 2d), indicating that caspase-12 negatively regulates or is dependent upon this pathway *in vivo*.

We next examined whether the catalytic function of rodent caspase-12 was required for its effects on inflammation. For this purpose, the human monocytic THP-1 cell-line—that does not produce human caspase-12 (our previous results⁵ and data not shown)—was transfected with rodent caspase-12 or the catalytically inactive form, caspase-12 Cys299Ala (along with green fluorescent protein (GFP) as a transfection marker), and cytokine production was examined in the sorted GFP-positive cells. Caspase-12 specifically inhibited the production of IL-1 β in response to TLR ligands (Fig. 3a) but did not have any effect on that of GM-CSF (Fig. 3b). Notably, mutation of the catalytic cysteine in caspase-12 did not abolish its inhibitory role on IL-1 β production, indicating that the enzymatic function of caspase-12 is not required for this effect. Addition of zVAD-fmk to block caspase-1 in the cultures resulted in undetectable levels of IL-1 β and IL-18 in the culture media but did not affect GM-CSF levels (data not shown). Therefore, in a manner similar to the way that cFLIP is a dominant-negative modulator of the 'extrinsic' cell death pathway^{9,10}, caspase-12 seems to be a counterpart in the inflammatory caspase cascade. This is supported by other observations, namely (1) the hyper-production of IL-1 β and IL-18 by *Casp12*^{-/-} splenocytes (Fig. 2b), (2) the inhibition of IL-1 β production by overexpressed rodent caspase-12 in THP-1 cells (Fig. 3a) (with the absence of an effect on pro-IL-1 β transcription (Fig. 2c)), and (3) the confinement of caspase-12 effects to caspase-1 substrates (for example, IL-1 β and IL-18 (substrates of caspase-1) but not GM-CSF).

Because caspase-12 seems to be a dominant-negative regulator of caspase-1, we tested this directly using two approaches. First, we examined the effect of rodent caspase-12 overexpression on the activity of caspase-1 in HEK 293T cells. Both wild-type caspase-12 and the catalytically inactive mutant caspase-12 Cys299Ala blocked the activity of caspase-1 on its preferred substrate Trp-Glu-His-Asp-7-amino-4-trifluoromethyl coumarin (WEHD-AFC) (Fig. 4a). Second, we used an established system in which caspase-1 is spontaneously activated in a cell-free system¹¹, in the presence of either a control protein or increasing concentrations of bacterially expressed and purified recombinant caspase-12, and examined the maturation of IL-1 β by western blot analysis (Fig. 4b). Caspase-12 inhibited the production of the 17-kDa mature form of IL-1 β in a dose-dependent manner, confirming that the dominant-negative inhibitory action of caspase-12 affected both synthetic and macromolecular substrates of caspase-1.

To determine whether the inhibitory role of caspase-12 on caspase-1 is mediated through an association between the two proteins, a co-immunoprecipitation experiment was performed in HEK 293T cells, revealing that caspase-12 co-immunoprecipitates with caspase-1. To control for specificity, we also examined the co-immunoprecipitation of caspase-12 with caspase-5 and caspase-9, both of which contain a CARD domain. Caspase-12 co-immunoprecipitated

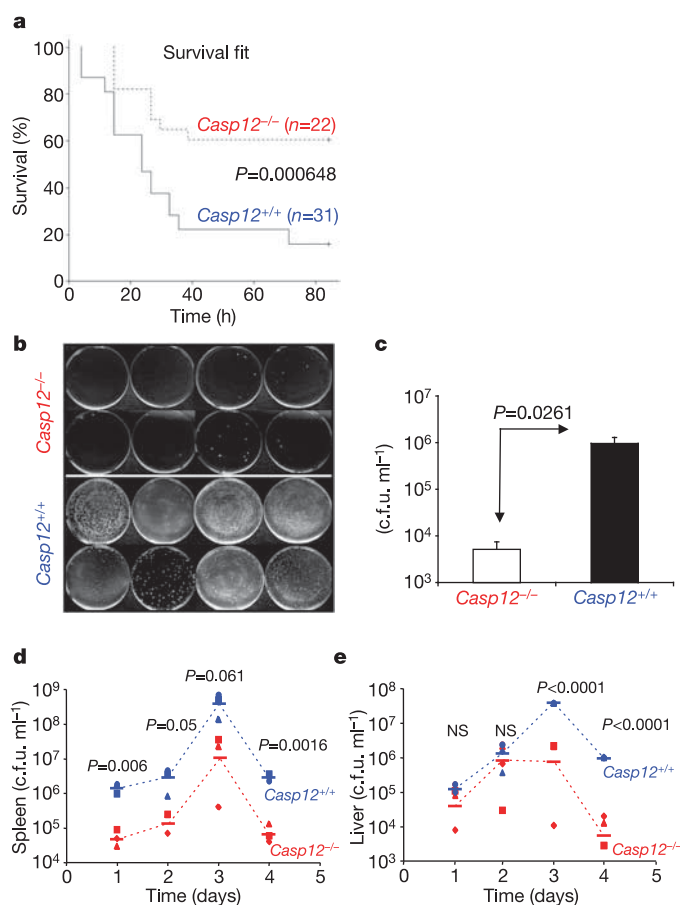


Figure 1 *Casp12*^{-/-} mice are resistant to sepsis and clear pathogens more efficiently than *Casp12*^{+/+} mice. **a**, Sepsis was induced using the CASP model of polymicrobial infection. Survival was recorded every 3 h. Whereas more than 80% of *Casp12*^{+/+} mice succumbed to septic shock in the first 48 h after CASP, most *Casp12*^{-/-} mice survived. **b**, Blood bacterial load in sepsis. Peripheral blood was collected 12 h after CASP, and serial dilutions were plated on blood agar plates, showing that *Casp12*^{-/-} mice cleared the polymicrobial infection more efficiently than wild-type littermates. Each plate represents a different mouse ($n = 8$ for both *Casp12*^{+/+} and *Casp12*^{-/-}). **c**, Quantification of c.f.u. per ml blood from the experiment described in **b**. Error bars indicate standard deviations. **d**, *Listeria monocytogenes* pathogen clearance. *Casp12*^{+/+} and *Casp12*^{-/-} mice were challenged intravenously with a defined dose of *L. monocytogenes*, and pathogen clearance was determined by assessment of c.f.u. ml⁻¹ of bacteria in spleen (**d**) and liver (**e**). NS, not significant. $n = 3$ per genotype per time point, indicated by individual symbols of the same colour.

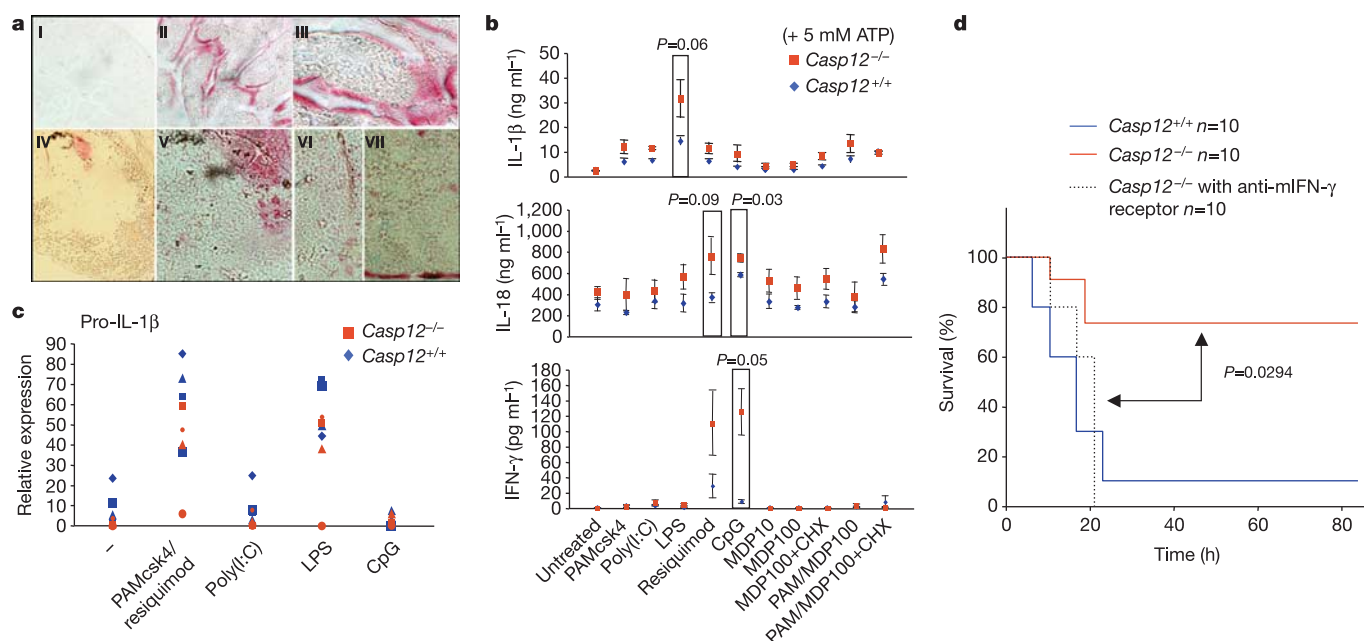


Figure 2 | Caspase-12 dampens production of cytokines essential for sepsis survival. **a**, Caspase-12 is expressed in intestinal epithelial cells and is induced in the spleen in response to bacteria. Caspase-12 expression was examined by staining tissue sections from *Casp12*^{-/-} mice that either underwent CASP (panels IV–VII) or not (panels II–III) for β-galactosidase expression (driven from the LacZ-neo cassette in the *Casp12* targeted allele). Panel I: intestine section from a wild-type mouse was stained for β-galactosidase as a negative control. **b**, Enhanced production of IL-1β, IL-18 and IFN-γ by *Casp12*^{-/-} splenocytes. Splenocytes from *Casp12*^{+/+} ($n = 4$) and *Casp12*^{-/-} ($n = 4$) mice were treated *ex vivo* with different TLR/NOD ligands for 24 h, and cytokine levels were measured from the

culture media, using bead-based multiplex immunoassays. Results are from two independent experiments. Error bars indicate standard deviations. **c**, Thioglycollate-elicited mouse peritoneal macrophages from *Casp12*^{+/+} ($n = 4$) and *Casp12*^{-/-} ($n = 4$) mice were treated with TLR ligands, mRNA extracted and pro-IL-1β levels measured by quantitative real-time PCR. Individual animals are represented by individual symbols of the same colour. **d**, Inhibition of IFN-γ signalling ablates the survival advantage of caspase-12-deficient mice after CASP. One hour before CASP, mice were injected intraperitoneally with 250 μg per mouse of neutralizing anti-mouse IFN-γ receptor antibodies or isotype control (not shown). Survival to sepsis was assessed as in Fig. 1a.

to a lesser extent with caspase-5 (compared with caspase-1) and did not co-immunoprecipitate with caspase-9 (Fig. 4c).

Whereas both caspase-1- and caspase-11-deficient mice are impaired in their ability to produce IL-1β and are resistant to lethal doses of endotoxin/LPS^{2,4}, caspase-1-deficient mice are two- to threefold more susceptible to lethal *Escherichia coli* infection than wild-type mice¹². Because caspase-12 is an inhibitor of caspase-1, the resistance to sepsis of *Casp12*^{-/-} mice is probably associated with an initial hyper-production of these cytokines mediated by a de-repression of caspase-1. The resulting beneficial effect of cytokine hyper-production is contrary to current dogma in sepsis research in which the initial cytokine 'storm' is viewed as harmful and has been the basis of many (albeit unsuccessful) clinical studies. Thus, in wild-type cells murine caspase-12 seems to attenuate the activity of caspase-1 that is normally essential for bacterial clearance and sepsis survival.

To evaluate the role of caspase-12 in the haematopoietic system and understand whether its expression in other tissues (for example, intestine) is also required for its effects on pathogen clearance, bone marrow chimaera studies were performed. *Casp12*^{+/+} or *Casp12*^{-/-} bone marrow was transplanted, respectively, into *Casp12*^{-/-} or *Casp12*^{+/+} recipient mice. The ability of the chimaeric mice to clear polymicrobial infection was determined as above in response to CASP-induced sepsis. Whereas *Casp12*^{+/+} bone marrow reduced the ability of *Casp12*^{-/-} mice to clear bacteria, the transfer of *Casp12*^{-/-} bone marrow into *Casp12*^{+/+} host mice only partially improved their clearance ability and did not rescue them from sepsis (Supplementary Fig. 4 and data not shown). These results suggest that caspase-12 functions not only in the haematopoietic system, but may also modulate the inflammatory response in other tissues.

It was recently reported that NOD2 modulates TLR2 signalling^{13,14}. We therefore examined whether caspase-12 has a role in

this modulation. Splenocytes from wild-type and *Casp12*^{-/-} mice were treated with the TLR2 ligand PAMcsk4 alone or in combination with the NOD2 ligand MDP, and the levels of IFN-γ and IL-6 were measured. The inhibitory effect of NOD2 signalling on IFN-γ production was not altered by the presence or absence of caspase-12 (ref. 12 and Supplementary Fig. 5a), indicating that caspase-12 does not cooperate with NOD2 to regulate IFN-γ levels. Similarly, the synergy between the TLR2 and NOD2 ligands in IL-6 production¹⁴ was not affected by the absence of caspase-12 (Supplementary Fig. 5b), suggesting that caspase-12 functions independently of NOD2 downstream of TLR signalling.

Murine caspase-12 was previously proposed as a key mediator of endoplasmic reticulum (ER) stress-induced apoptosis¹⁵. Because apoptosis contributes adversely to sepsis progression¹, we sought to evaluate the role of caspase-12 in ER stress-induced apoptosis in our system. We examined the response of primary mouse embryonic fibroblasts (MEFs)—isolated from wild-type and *Casp12*^{-/-} mice—to apoptosis induced by the ER stressors brefeldin A, tunicamycin, thapsigargin and the calcium ionophore A23187. *Casp12*^{+/+} and *Casp12*^{-/-} MEFs showed similar levels of cell death after treatment with these stressors (Supplementary Fig. 6). Similarly, the presence or absence of caspase-12 had no effect on apoptotic sensitivity to diverse stimuli, including activators of the extrinsic and intrinsic cell death pathways (data not shown). These data are consistent with those described for human caspase-12 in which no differential apoptosis sensitivity was observed⁵, as well as with other results^{16,17}; however, they differ from what was reported in another study¹⁵. Because the mice described here were generated using a different targeting strategy than that of the mice described in the latter study¹⁵, the difference in results may relate to a difference in caspase-11 expression levels in those animals versus those we used. Caspase-11 is not only an important player in responses to endotoxin⁴, but it has

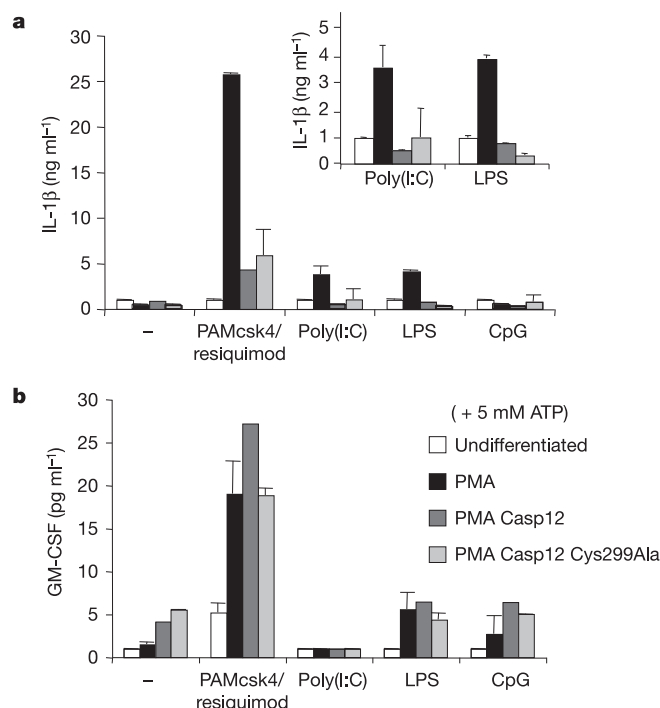


Figure 3 | The catalytic function of caspase-12 is not required for its inhibitory effect on IL-1 β production. **a**, **b**, THP-1 cells (deficient in human caspase-12 expression) were reconstituted to express GFP in combination with vector control, rat caspase-12 or the catalytically inactive mutant Cys299Ala. Sorted live GFP-positive cells were either left undifferentiated or were differentiated with PMA before treatment with TLR ligands. Levels of IL-1 β (**a**) and GM-CSF (**b**) were measured from the culture media using bead-based human immunoassays. Treatments were done in triplicate. Results are from two independent experiments. Error bars indicate standard deviations.

also been suggested to have a role in ER stress apoptosis¹⁸. It is therefore possible that caspase-11 rather than caspase-12 might have a role in ER stress-induced apoptosis in some settings.

Our study describes a fundamentally new role for a caspase: dampening of responses to bacterial infection. Murine caspase-12 deficiency confers resistance to sepsis and its presence exerts a direct suppressive effect on caspase-1, resulting in enhanced vulnerability to bacterial infection and septic shock. Whereas the presence of caspase-12 seems to be detrimental for the individual during sepsis, in both mice (this study) and humans⁵, it remains to be determined whether murine caspase-12 and the human CASP12L variant modulate inflammation and bacterial clearance by overlapping or mutually exclusive mechanisms. Regardless, caspase-12 is detrimental to *in vivo* handling of systemic bacterial infections and predisposes to sepsis, thereby making it a potentially important target for future therapeutic strategies.

METHODS

Detailed methods and methods not described here can be found in Supplementary Information.

Generation of Casp12 knockout mice and primary cell cultures. *Casp12* knockout mice were generated by Deltagen under a consortium agreement involving Merck & Co. A targeting construct was engineered to disrupt the *Casp12* gene in ES cells by homologous recombination. A segment of exon 2 was targeted for replacement by a cassette consisting of the β -galactosidase gene and the neomycin (*neo*) resistance gene. The *Casp12*^{-/-} mice used in this study were backcrossed to homogeneity to a C57BL/6J background. MEFs were prepared from embryos harvested at day 14 of gestation. Mouse intestinal epithelial cells (IECs) were purified on discontinuous Percoll gradients from the small intestine. **Induction of sepsis using the CASP procedure.** Sepsis was induced in mice aged between 6–12 weeks by pressing caecal content into the abdomen through a stent

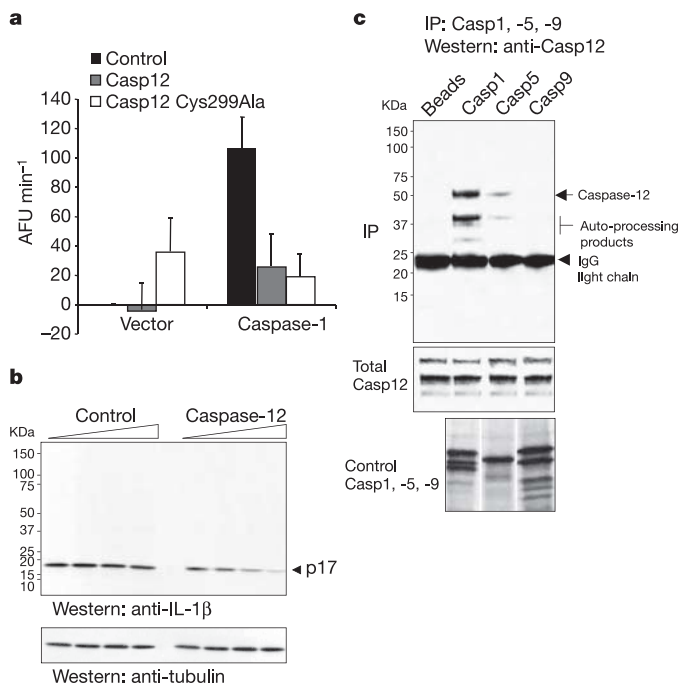


Figure 4 | Caspase-12 as well as the catalytically inactive mutant Cys299Ala associate with caspase-1 and block its activity. **a**, Caspase-1 substrate WEHD-AFC cleavage by HEK 293T lysates expressing either vector control or caspase-1 in combination with caspase-12 or caspase-12 Cys299Ala. Error bars indicate standard deviations. AFU, arbitrary fluorescent units. **b**, Caspase-12 inhibits inflammasome-mediated IL-1 β maturation in a dose-dependent manner. THP-1 lysates from LPS-prestimulated cells were incubated at 30°C in the presence of either GST (control) or GST-rat caspase-12, and IL-1 β maturation was detected by western blot analysis using an antibody against the mature form of IL-1 β . **c**, Caspase-12 co-immunoprecipitates with caspase-1. HEK 293T cells were co-transfected with Flag-tagged caspase-1, caspase-5 or caspase-9 and caspase-12. Flag-tagged caspases were immunoprecipitated from HEK 293T cells with Flag M2 agarose beads and the association of caspase-12 was examined by western blot of the immunoprecipitated product using anti-mouse caspase-12 antibodies. 'Total' represents 10% of the total cell lysates used in the immunoprecipitation (IP) reaction. The caspase-12 processing products probably result from autocatalysis as they were not present when the catalytically inactive caspase-12 C299A mutant was transfected in 293T cells (not shown).

positioned in the caecal wall. After surgery, mice were monitored five times per day during the first 48 h after surgery, observing activity, alertness and temperature. Animals that were determined to be moribund (temperature <23°C, showing minimal response to challenge and diminished righting reflex) were killed, allowing for collection of blood and for post-mortem analysis. Long-term survivors were followed until day 8 after surgery when they were killed for final sampling and necropsy.

Listeria monocytogenes challenge. For *in vivo* infections, *Listeria monocytogenes* strain DP-L4056 (a gift from D. Portnoy) was grown to mid-logarithmic phase. Approximately 5×10^4 c.f.u. were injected into the lateral tail vein.

Determination of bacterial content in sepsis and Listeria models. In the sepsis model, 100 μ l peripheral blood samples were collected by orbital puncture 12 h after CASP. The blood volume was brought up to 0.5 ml with distilled water and serial dilutions of that (1:2, 1:100, 1:1,000) were prepared in distilled water and were plated on brain heart infusion blood agar plates. After a 24-h incubation of the plates, bacterial colonies were counted and the number of c.f.u. per ml blood was calculated. In the *Listeria* model, peripheral blood was collected and analysed as above on days 1, 2, 3 and 4 after *Listeria* injection. At the same time points, spleens and livers were also collected and were homogenized in a 10-ml volume of a 0.2% NP-40 solution. Serial dilutions (1:10, 1:100, 1:1,000) were plated on brain heart infusion agar and bacterial counts were assessed as above.

Splenocyte culture and treatments. Splenocytes were isolated by mechanical disruption followed by differential centrifugation and re-suspension in RPMI.

Cells were treated for 24 h with PAMcsk4 (100 ng ml^{-1} , TLR2), poly(I:C) ($4 \mu\text{g ml}^{-1}$, TLR3), LPS Re595 (*Salmonella minnesota*; 100 ng ml^{-1} , TLR4), resiquimod R848 ($1:500$, TLR7), CpG ($0.2 \mu\text{M}$, TLR9), and muramyl dipeptide MDP ($10 \mu\text{M}$ or $100 \mu\text{M}$) with or without cycloheximide CHX ($0.5 \mu\text{g ml}^{-1}$). Cells were then treated with 5 mM ATP in fresh media for 20 min, washed and further cultured in fresh media for 3 h. A 10-plex (ten mouse cytokines: IL-1 β , IL-2, IL-4, IL-5, IL-6, IL-10, IL-12p70, IFN- γ , TNF- α and GM-CSF) bead-based immunoassay kit from LINCOplex was used according to the manufacture's instructions. Data were obtained and analysed using the Bio-Plex system from Biorad.

Anti-IFN- γ receptor treatment. Anti-IFN- γ antibodies and the isotype control antibodies were purchased from PharMingen (catalogue number 557530 and 553969, respectively). The antibodies were administered via intraperitoneal injection at a dose of $250 \mu\text{g}$ per mouse 1 h before induction of sepsis by CASP.

THP-1 cell transfection and differentiation. THP-1 cells were transfected by electroporation with $20 \mu\text{g}$ DNA ($15 \mu\text{g}$ caspase-12 (either wild type or Cys299Ala mutant) plus $5 \mu\text{g}$ GFP). After electroporation, cells were kept on ice for 5 min then grown in 20 ml media for 24 h. Four electroporation reactions (that is, 80 million cells) were pooled for each caspase-12 (either wild type or Cys299Ala mutant) transfection. The next day, live GFP-positive cells were sorted, pre-treated with 100 ng ml^{-1} phorbol myristate acetate (PMA) for 6 h then treated with TLR ligands for 24 h. For IL-1 β measurement, cells were then treated with 5 mM ATP in fresh media for 20 min, washed and further cultured in fresh media for 3 h. Cell supernatants were then analysed for cytokine production using bead-based multiplex human immunoassays (described above).

Caspase-1 activity assays. HEK 293T cells were co-transfected using lipofectamine plus reagent (Invitrogen) with Flag-tagged caspase-1 and caspase-12 (either wild type or Cys299Ala mutant). Twenty-four hours after transfection, cell lysates were prepared in caspase activation buffer (20 mM PIPES pH 7.4, 100 mM NaCl, 1 mM EDTA, 10 mM dithiothreitol (added fresh), 0.1 CHAPS, 10% sucrose and protease inhibitors), and $100 \mu\text{g}$ lysates were assayed for cleavage of $100 \mu\text{M}$ of the caspase-1 substrate WEHD-AFC. Kinetic measurements of AFC release were recorded every minute for 30 min at room temperature using a fluorescent plate reader (excitation 400 nm, emission 505 nm).

Cell-free inflammasome activation assay. As in ref. 11, THP-1 cells were pre-stimulated with $1 \mu\text{g ml}^{-1}$ LPS for 1 h. Cells were harvested and washed with PBS, then lysed by hypo-osmotic swelling followed by mechanical disruption by passage 15 times through a 22-gauge needle. Cell lysates were centrifuged, and the supernatants, after filtration ($0.45 \mu\text{M}$), were used for this *in vitro* IL-1 β cleavage assay. The lysates were incubated for 20 min on ice with either GST (control) or GST-rat caspase-12, then shifted to 30°C and incubated for 60 min. IL-1 β processing was monitored by western analysis with a specific antibody against the mature form of IL-1 β (Cell Signaling, catalogue number 2021).

Co-immunoprecipitation experiments. HEK 293T cells were co-transfected using lipofectamine plus reagent (Invitrogen) with Flag-tagged caspase-1, caspase-5, caspase-9 (Cys-to-Ala mutants) and caspase-12. Twenty-four hours after transfection, cell lysates were obtained by detergent disruption and caspase-1, caspase-5 and caspase-9 were immunoharvested using Flag M2 agarose beads (Sigma). Immunocomplexes were eluted from the beads using Flag peptides (Sigma) and were processed for western blot analysis using anti-mouse caspase-12 antibodies or anti-Flag antibodies (Sigma).

Other methods. Cell death was triggered in MEFs and was measured by flow cytometry after propidium iodide staining. Immunohistochemical procedures were performed on sections fixed in 1% formaldehyde. Bone marrow chimaeric mice were generated after transfer of harvested bone marrow into recipient mice that had been ablated with two doses of irradiation at 550 rad. NF- κB activity was determined in cells that had been transfected with κB -luciferase and β -gal reporter plasmids. Detailed descriptions of these and all methods are available in Supplementary Information.

Statistical analysis. Log rank test and Fisher's exact test were used to compare sepsis survival among *Casp12*^{+/+} and *Casp12*^{-/-} mice. Student *t*-test was used

for statistical analyses of apoptosis, bacterial clearance as well as cytokine production assays.

Received 6 December 2005; accepted 14 February 2006.

- Hotchkiss, R. S. *et al.* Caspase inhibitors improve survival in sepsis: a critical role of the lymphocyte. *Nature Immunol.* **1**, 496–501 (2000).
- Li, P. *et al.* Mice deficient in IL-1 β -converting enzyme are defective in production of mature IL-1 β and resistant to endotoxin shock. *Cell* **80**, 401–411 (1995).
- Kuida, K. *et al.* Altered cytokine export and apoptosis in mice deficient in interleukin-1 β converting enzyme. *Science* **267**, 2000–2003 (1995).
- Wang, S. *et al.* Murine caspase-11, an ICE-interacting protease, is essential for the activation of ICE. *Cell* **92**, 501–509 (1998).
- Saleh, M. *et al.* Differential modulation of endotoxin responsiveness by human caspase-12 polymorphisms. *Nature* **429**, 75–79 (2004).
- Zantl, N. *et al.* Essential role of gamma interferon in survival of colon ascendens stent peritonitis, a novel murine model of abdominal sepsis. *Infect. Immun.* **66**, 2300–2309 (1998).
- Hotchkiss, R. S. *et al.* Adoptive transfer of apoptotic splenocytes worsens survival, whereas adoptive transfer of necrotic splenocytes improves survival in sepsis. *Proc. Natl Acad. Sci. USA* **100**, 6724–6729 (2003).
- Prass, K. *et al.* Stroke-induced immunodeficiency promotes spontaneous bacterial infections and is mediated by sympathetic activation reversal by poststroke T helper cell type 1-like immunostimulation. *J. Exp. Med.* **198**, 725–736 (2003).
- Rasper, D. M. *et al.* Cell death attenuation by 'Usurpin', a mammalian DED-caspase homologue that precludes caspase-8 recruitment and activation by the CD-95 (Fas, APO-1) receptor complex. *Cell Death Differ.* **5**, 271–288 (1998).
- Irmeler, M. *et al.* Inhibition of death receptor signals by cellular FLIP. *Nature* **388**, 190–195 (1997).
- Martinon, F., Burns, K. & Tschopp, J. The inflammasome: a molecular platform triggering activation of inflammatory caspases and processing of proIL- β . *Mol. Cell* **10**, 417–426 (2002).
- Joshi, V. D., Kalvakolanu, D. V., Hebel, J. R., Hasday, J. D. & Cross, A. S. Role of caspase 1 in murine antibacterial host defenses and lethal endotoxemia. *Infect. Immun.* **70**, 6896–6903 (2002).
- Watanabe, T., Kitani, A., Murray, P. J. & Strober, W. NOD2 is a negative regulator of Toll-like receptor 2-mediated T helper type 1 responses. *Nature Immunol.* **5**, 800–808 (2004).
- Kobayashi, K. S. *et al.* Nod2-dependent regulation of innate and adaptive immunity in the intestinal tract. *Science* **307**, 731–734 (2005).
- Nakagawa, T. *et al.* Caspase-12 mediates endoplasmic reticulum-specific apoptosis and cytotoxicity by amyloid- β . *Nature* **403**, 98–103 (2000).
- Kalai, M. *et al.* Regulation of the expression and processing of caspase-12. *J. Cell Biol.* **162**, 457–467 (2003).
- Obeng, E. A. & Boise, L. H. Caspase-12 and caspase-4 are not required for caspase-dependent endoplasmic reticulum stress-induced apoptosis. *J. Biol. Chem.* **280**, 29578–29587 (2005).
- Hitomi, J. *et al.* Involvement of caspase-4 in endoplasmic reticulum stress-induced apoptosis and A β -induced cell death. *J. Cell Biol.* **165**, 347–356 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements M.S. is supported by a CIHR post-doctoral fellowship. D.R.G. is supported by grants from the US NIH. We thank S. Granger and A. Coddington for help with the Bio-plex system and C. Bonzon for help with MEF preparation.

Author Contributions D.R.G. and D.W.N. share senior authorship.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper on www.nature.com. Correspondence and requests for materials should be addressed to D.W.N. (donauld_nicholson@merck.com) or D.R.G. (douglas.green@stjude.org).

Spatiotemporal dynamics of RhoA activity in migrating cells

Olivier Pertz^{1,2†}, Louis Hodgson^{1,2†}, Richard L. Klemke^{2†} & Klaus M. Hahn^{1,2†}

Rho family GTPases regulate the actin and adhesion dynamics that control cell migration. Current models postulate that Rac promotes membrane protrusion at the leading edge and that RhoA regulates contractility in the cell body^{1,2}. However, there is evidence that RhoA also regulates membrane protrusion^{3,4}. Here we use a fluorescent biosensor, based on a novel design preserving reversible membrane interactions, to visualize the spatiotemporal dynamics of RhoA activity during cell migration. In randomly migrating cells, RhoA activity is concentrated in a sharp band directly at the edge of protrusions. It is observed sporadically in retracting tails, and is low in the cell body. RhoA activity is also associated with peripheral ruffles and pinocytic vesicles, but not with dorsal ruffles induced by platelet-derived growth factor (PDGF). In contrast to randomly migrating cells, PDGF-induced membrane protrusions have low RhoA activity, potentially because PDGF strongly activates Rac, which has previously been shown to antagonize RhoA activity^{5,6}. Our data therefore show that different extracellular cues induce distinct patterns of RhoA signalling during membrane protrusion.

To study the spatiotemporal dynamics of RhoA activation in living cells, we designed a genetically encoded, single-chain biosensor with intramolecular fluorescence resonance energy transfer (FRET) that responds to RhoA activation. As depicted in Fig. 1a, the biosensor consists of a Rho-binding domain of the effector rhotein (RBD, amino acids 7–89), which specifically binds to GTP-RhoA⁷, followed by cyan fluorescent protein (CFP), an unstructured linker of optimized length, a pH-insensitive variant of yellow fluorescent protein⁸ (YFP), and full-length RhoA. Upon activation by GTP-loading, the RBD binds to Rho, modifying the relative orientation of the two fluorophores and thereby increasing FRET. Because the fluorescent proteins are attached to one another, RhoA activation can be approximated simply as being proportional to the FRET/CFP emission ratio at a given subcellular location⁹. The two fluorophores are placed on the internal portion of the biosensor chain, leaving the carboxy terminus of RhoA intact for binding to Rho guanine dissociation inhibitor¹⁰ (RhoGDI). This important regulatory interaction controls reversible membrane localization.

As the size of the biosensor precluded purification for *in vitro* characterization, we analysed the fluorescence spectra of mutant biosensors expressed in suspensions of living HEK-293T cells. G14V and Q63L dominant-positive mutants showed higher emission ratios than a T19N dominant-negative or an F39A effector binding mutant unable to interact with the RBD⁷ (Fig. 1b, c). Notably, the emission ratios of the wild-type and dominant-positive biosensors were similar (Fig. 1b, c). This is due to the high biosensor expression level required to obtain a reasonable signal-to-noise ratio in this assay (more than an order of magnitude above endogenous RhoA levels,

Supplementary Fig. S1a). At this concentration, there is insufficient endogenous RhoGDI to prevent the biosensor from accumulating at the plasma membrane (Supplementary Fig. S1b), where it can be activated by guanine nucleotide exchange factors (GEFs). Emission ratios of wild-type biosensor were reduced by co-expression of RhoGDI, p50 Rho GTPase activating protein (GAP) or C3-transferase, but not by Rap GAP, which does not interact with RhoA (Fig. 1c). To test the ability of the biosensor to respond to GEFs, we first sequestered it in the cytosol by co-expressing RhoGDI. Co-expression of truncated, constitutively active GEFs that can activate RhoA (Dbl and Ect2) rescued the high emission ratio and membrane localization, whereas GEFs that do not interact with RhoA (Tiam1

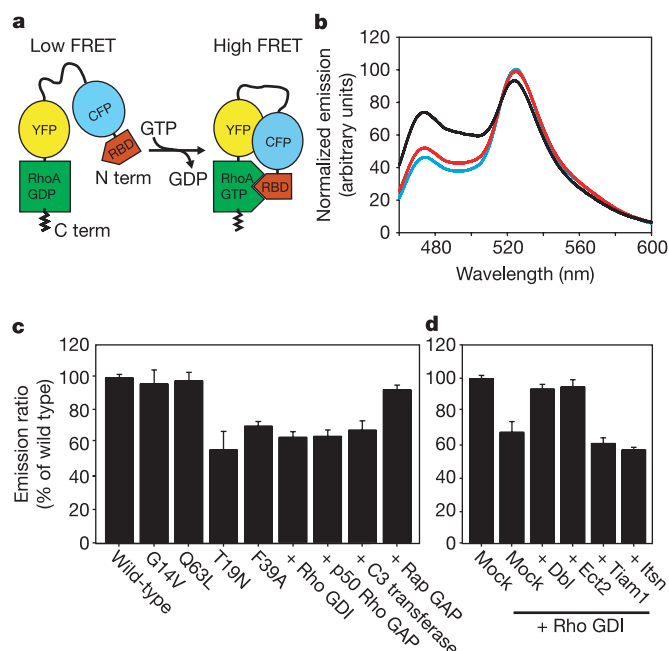


Figure 1 | Design and characterization of the RhoA biosensor. **a**, Probe design. **b**, Normalized emission spectra of biosensor mutants expressed in HEK-293T cells (excitation at 433 nm). Blue, wild-type biosensor; red, Q63L dominant-positive biosensor; black, T19N dominant-negative biosensor. **c**, **d**, Normalized FRET/CFP emission ratios: mutant RhoA biosensors and wild-type biosensor co-expressed with negative regulators (**c**) or GEFs (**d**) as shown. Results are the mean of at least three independent measurements. Error bars represent standard deviations; $P < 0.001$ (*t*-test) for differences between wild type and signals from T19N, F39A, +Rho GDI, +p50 Rho GAP, +C3 transferase (in **c**) and +Rho GDI, +Tiam1, +Its1n (in **d**).

¹University of North Carolina at Chapel Hill, Department of Pharmacology and Lineberger Cancer Center, Chapel Hill, North Carolina 27599, USA. ²Department of Cell Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA. [†]Present addresses: University of California at San Diego, Department of Pathology and Moores Cancer Center, Basic Science Building, Room 1040, 9500 Gilman Drive, MC 0612, La Jolla, California 92093, USA (O.P., R.L.K.); University of North Carolina at Chapel Hill, Department of Pharmacology and Lineberger Cancer Center, Chapel Hill, North Carolina 27599, USA (L.H., K.M.H.).

and intersectin) had no effect (Fig. 1d and Supplementary Fig. S1b). Efficient co-expression of these constructs was verified by western blotting (Supplementary Fig. S1c).

For live-cell imaging, we developed a tetracycline-inducible mouse embryonic fibroblast (MEF) cell line that enabled biosensor expression at levels similar to endogenous RhoA, at which RhoGDI binding capacity is not overwhelmed and cell morphology and motile behaviour are not altered (Supplementary Movie S1). Further characterization of the biosensor in the MEF and HEK-293T cell systems is presented in the Supplementary Information. In non-motile cells (plated on fibronectin overnight), RhoA activity was always much higher at the periphery than in the cell body (Supplementary Fig. S5b). We observed a previously undescribed perinuclear accumulation of RhoA that colocalizes with a Golgi marker (Supplementary Fig. S5b). This pool of RhoA consistently showed among the lowest activation values in the cell. The significance of an inactive, Golgi-resident pool of RhoA is unclear. However, the Golgi-resident protein ARAP1 (ref. 11) has GAP activity for RhoA and may inactivate it at this location. ARAP1 also regulates Arf1, suggesting that this compartment may serve as a platform for the previously documented crosstalk between RhoA and Arf1¹².

We next examined the localization of RhoA activation in MEF cells randomly migrating on fibronectin (optimal motile behaviour was observed six hours after re-plating). Activation was consistently observed wherever protrusions were extending (Fig. 2a, c and Supplementary Movies S2, S3). There was a sharp band of markedly higher activity immediately adjacent to the cell edge (Fig. 2b, d), superimposed on a broader gradient of lower activation decreasing from the edge of the cell inwards. This was observed in 17 out of 19 protrusions ($n = 10$ cells), and was also found in cells migrating out of a wounded monolayer (data not shown). RhoA activity was infrequently observed during tail retraction. However, activation was obvious when such contractile behaviour was very robust (Fig. 2f, g; $n = 5$ cells). Only minimal RhoA activity was found in the cell body.

As an independent measure of RhoA activation in cell protrusions, we isolated pseudopods extending towards a gradient of fibronectin using a recently described technique¹³. Rhotekin pull-down assays showed that RhoA activity was higher in pseudopods than in the cell body (Fig. 2e). In this assay, Erk2 has been shown to be equally distributed in the pseudopod and body fractions, and serves as a loading control. Further pointing to a role for RhoA in membrane protrusion, low-level expression of dominant-negative T19N GFP-RhoA or T19N biosensor produced cells with concave edges that were unable to produce lamellae (Supplementary Fig. S3a). High expression of this mutant led to cell detachment (data not shown). Previous studies are consistent with roles for RhoA effectors directly at the leading edge, including adducin phosphorylation by Rho kinase³ (ROK), stabilization of microtubules by diaphanous-related formin (mDia)⁴, and potentially ROK-dependent myosin phosphorylation in a broader region throughout the lamellipod¹⁴. At the cell rear, RhoA might interact with ROK to turn on myosin¹⁵ during robust contractile events. The absence of high RhoA activity in the cell body challenges the notion that RhoA regulates focal adhesion assembly in the cell body². However, the conclusion in ref. 2 was reached using tools (C3-transferase, dominant-negative and dominant-positive constructs) that do not discriminate between different Rho family isoforms (RhoA, RhoB, RhoC) or their downstream signalling. It is therefore possible that some functions attributed to RhoA are due to different Rho isoforms.

We also detected high RhoA activity at the distal side of serum-induced peripheral ruffles (Fig. 3a, Supplementary Fig. S6a, b and Supplementary Movie S4; observed in 30 out of 38 ruffling events, $n = 8$ cells). RhoA activity trailed ruffles as they moved rearwards on the dorsal side of the cell, and ceased as the ruffle disappeared. This was not an artefact of cell motion between acquisition of the FRET and CFP images, as the same result was obtained with fixed cells or when the images were acquired in reverse order (data not shown). Neither was this FRET signal an artefact of ruffle thickness (Supplementary Fig. S6c), as use of a dominant-negative biosensor greatly

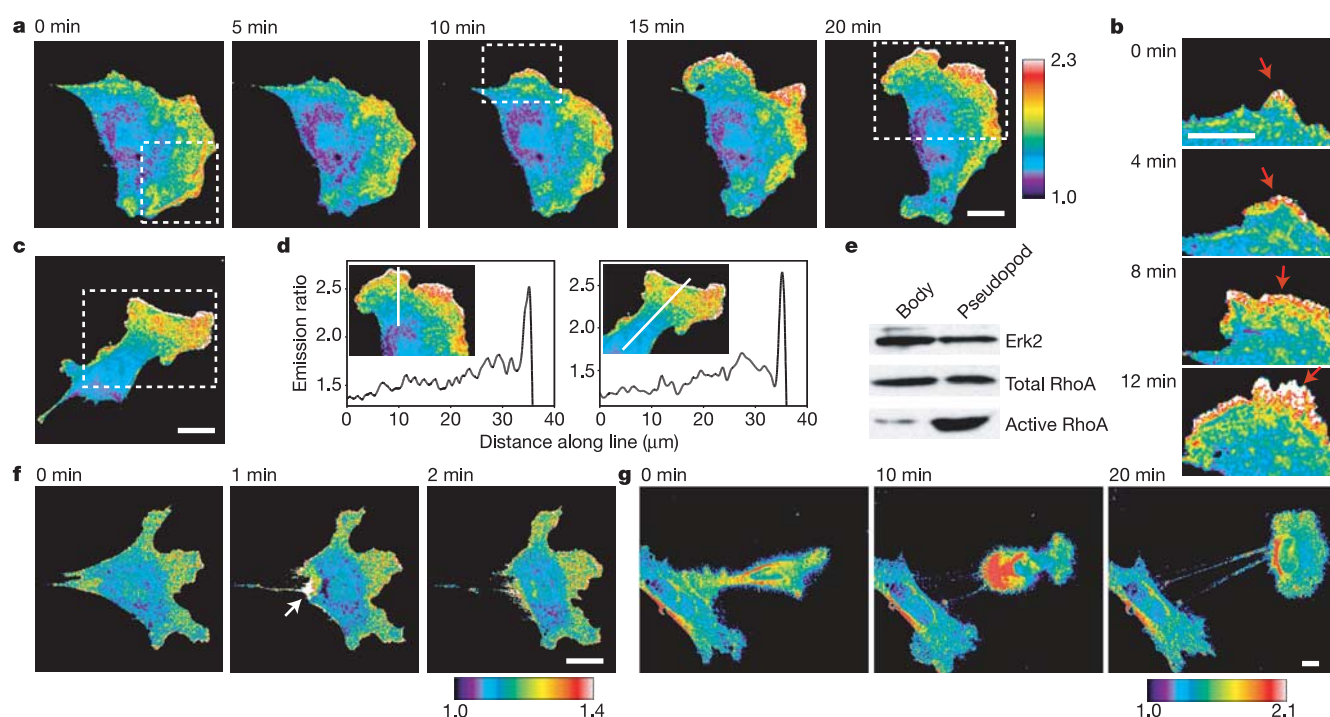


Figure 2 | RhoA activity in randomly migrating MEFs. **a, c,** MEFs migrating on fibronectin. **b,** Dynamics of a single protrusion from the region outlined in the $t = 10$ min frame of **a**. **d,** Emission ratio profiles along the indicated lines from insets at $t = 20$ min in **a** and in **c**. **e,** Western blot analysis of

endogenous RhoA activity in purified pseudopod and cell body fractions from cells migrating in a fibronectin gradient. **f, g,** RhoA activation patterns during tail retraction. Images in **a, c** and **f** are scaled so that regions of intense RhoA activity are shown in white. Scale bars, 20 μ m.

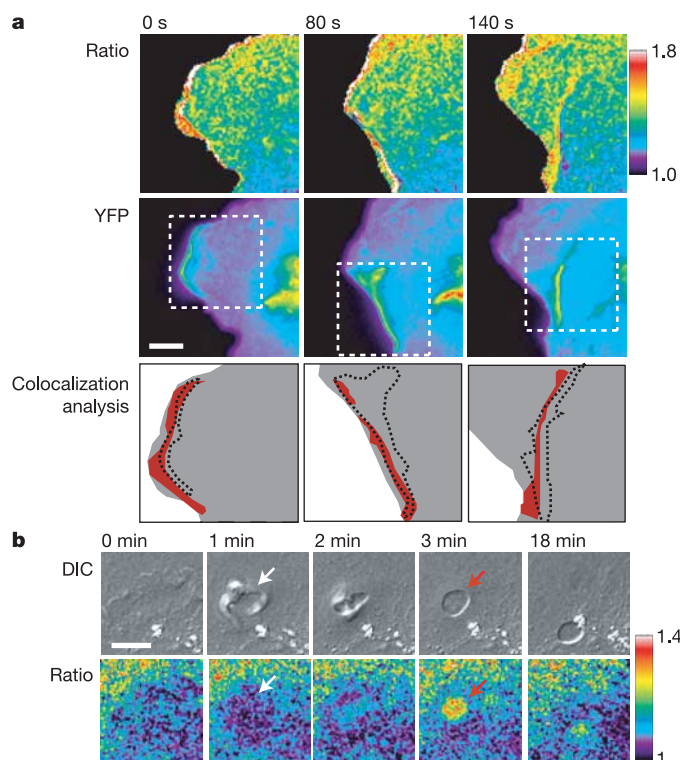


Figure 3 | RhoA activity at ruffles and pinosomes. **a**, Time-lapse images of peripheral ruffle movement induced by 10% serum. Zones of high YFP fluorescence mark ruffle location (see Supplementary Information). In the colocalization panels (from insets in the YFP images), red represents high RhoA activation and the dotted line shows ruffle position. **b**, Time-lapse images of PDGF-induced pinocytic vesicle formation. White arrows point to the absence of high RhoA activity at crown-like dorsal ruffles before pinocytic vesicle formation, red arrows point to high RhoA activity in mature pinocytic vesicles. Scale bars, 10 μ m.

reduced the emission ratio associated with the ruffles (Supplementary Fig. S6d). For reasons explained in the Supplementary Information, this activation pattern suggests that RhoA is activated throughout the distal, but not the proximal, membrane sheet of the ruffle. A similar activation pattern has also been observed using a different RhoA FRET probe¹⁶. This polarized RhoA activity might be important for ruffle movement, and may involve RhoA-dependent relocation of the ezrin-radixin-moesin proteins¹⁷ or mDia¹⁶ to this location.

PDGF also stimulated peripheral ruffling with identical RhoA activation patterns (data not shown). In contrast, there was no RhoA activation at dorsal, crown-like ruffles produced during PDGF-induced formation of pinosomes (Fig. 3b and Supplementary Movie S5; 30 out of 30 crown ruffles observed, $n = 18$ cells). This implies that distinct signalling mechanisms regulate actin dynamics at different ruffling structures. RhoA was suddenly activated at the pinosomes as they formed, and remained active as the pinosomes trafficked through cells (Fig. 3b; 14 out of 23 vesicles, $n = 18$ cells). Similar to the recent observation of mDia1-dependent regulation of actin coat formation around endosomes downstream of RhoB¹⁸, RhoA and mDia1 might regulate the bursts of actin polymerization that enable pinosome propulsion in the cytosol¹⁹.

PDGF stimulation induced strong lamellipodial protrusions in non-motile cells (re-plated overnight in the presence of 2% serum). However, unlike the randomly migrating cells described above, these protrusions lacked intense RhoA activity at the cell edges, and showed lower RhoA activity throughout the protrusion (Fig. 4a, b and Supplementary Movie S6; observed in all of 41 protrusions, $n = 19$ cells). To support these results, we isolated pseudopods of MEFs migrating towards a gradient of PDGF, and also observed low levels of endogenous RhoA activity in the pseudopods relative to the cell body (Fig. 4c). However, in cells with impaired RhoA function (for example, expressing a dominant-negative biosensor), PDGF stimulation only partially rescued protrusion (Supplementary Movie S7), suggesting that some RhoA activity is important for PDGF-induced membrane extension. PDGF potently activates Rac²⁰, leading to downregulation of RhoA activity in fibroblasts⁶. We therefore tested whether active Rac could turn off RhoA activity in membrane protrusions. Expression of dominant-positive Rac Q61L tagged with monomeric red fluorescent protein (mRFP) led to the

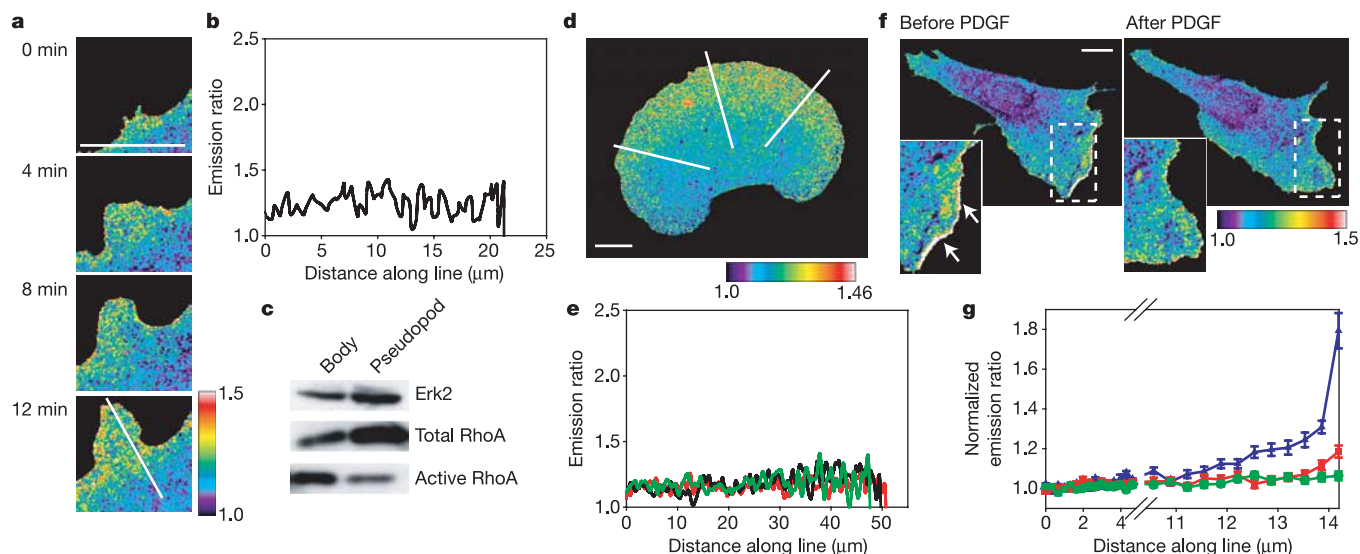


Figure 4 | PDGF-induced membrane protrusions. **a**, Dynamics of a PDGF-induced protrusion. **b**, Emission ratio profile along the line in the 12 min panel in **a**. **c**, Western blot analysis of endogenous RhoA activity in pseudopod and body fractions from cells migrating to a PDGF gradient. **d**, RhoA activity in a cell expressing dominant-positive mRFP-Rac Q61L.

e, Emission ratio profiles along the lines in **d**. **f**, Cells migrating out of a wounded monolayer before and after treatment with PDGF. **g**, Averaged emission ratio profiles from multiple cells as in **d** (green, 15 cells, 84 lines) and **f** (blue, 7 min before PDGF, 14 cells, 43 lines; red, 7 min after PDGF, 13 cells, 45 lines). Error bars show s.d. Scale bars, 30 μ m.

well-documented protrusive phenotype while recapitulating the RhoA activation pattern seen in PDGF-induced protrusion (Fig. 4d, e, g). This was also observed with a dominant-positive Rac G12V mutant (data not shown). PDGF stimulation also switched off pre-existing intense RhoA activation at the edges of MEFs migrating out of a wounded monolayer (Fig. 4f, g and Supplementary Movie S8). This data shows that PDGF stimulation and Rac activation downregulate RhoA activity specifically at the leading edge of migrating cells.

Our results show that RhoA activity is different in membrane protrusions induced by growth factor stimulation compared to that observed during random cell migration or in cells in a wounded monolayer. Antagonism between Rac and RhoA^{5,6} has a prominent role at the leading edge during growth factor stimulation. These multiple mechanisms regulating membrane protrusion may provide the means to produce different modes of cell motility and polarization in response to different extracellular cues.

METHODS

Microscopy and image acquisition. MEF/3T3 fibroblasts were plated on fibronectin-coated glass coverslips (10 µg ml⁻¹ fibronectin for 1 h at 37 °C, Sigma). Three-to-four hours after plating, cells were imaged in Ham's F-12K medium without phenol red (Biosource) with 10% fetal bovine serum, 25 mM HEPES and 10 µg ml⁻¹ oxyrase reagent (Oxyrase Inc.), in a heated closed chamber. Images were obtained using a Zeiss ×40 1.3 NA Plan NeoFluar DIC lens, a Zeiss Axiovert 100TV microscope, a Quantix CCD camera (Roper Scientific) and Metamorph software (Universal Imaging).

For emission ratio imaging, the following filter sets were used (Chroma): CFP: D436/20, D470/40; FRET: D436/20, HQ535/30; YFP: HQ500/20, HQ535/30. mRFP was imaged using HQ580/30 and HQ630/40. A dichroic mirror was custom-manufactured by Chroma for compatibility with all of these filters. Cells were illuminated with a 200 W Hg lamp through a 36% neutral density filter. At each time point, three images were recorded with the following exposure times: CFP (1 s), FRET (0.5 s), YFP (0.5 s) at binning 2 × 2 or 3 × 3. In some experiments, additional differential interference contrast (DIC) and/or mRFP pictures were also recorded.

For PDGF perfusion experiments (PDGF-BB from Sigma), a heated, open perfusion chamber (Harvard Apparatus) was used to maintain cells at 37 °C on the stage. The perfusion system consisted of a digitally controlled valve assembly regulating the perfusate, driven by gravity flow. The perfusate influx line was heated using a solution heater (Harvard Apparatus) to minimize temperature drift. Throughout the experiment, fluid flow rate was maintained at approximately 1 ml min⁻¹, preventing any effects due to sudden changes in flow rate, as demonstrated using vehicle controls. For wound-healing studies, MEFs were mixed at a 1:5 concentration ratio with MEF/3T3 cells repressed with doxycycline, plated on fibronectin-coated coverslips at 1 × 10⁶ cells per coverslip, and allowed to adhere for 4 h. Confluent monolayers were then wounded using a razor blade, and allowed to recover for an additional 4–5 h before observation.

Image processing. Metamorph software was used to perform image analysis. All images were first shading-corrected and background-subtracted. The FRET image, because it had the largest signal-to-noise ratio and therefore provided the best distinction between the cell and the background, was thresholded to generate a binary mask with a value of zero outside the cell and a value of one inside the cell. After multiplication by this mask, the FRET image was divided by the CFP image to yield a ratio image reflecting RhoA activation throughout the cell. A linear pseudocolour lookup table was applied, and the ratio values were normalized to the lower scale value. In each experiment, CFP and FRET images were carefully inspected to verify that all portions used to create the ratio image had a high enough signal-to-noise ratio. This was especially important in thin parts of the cell where fluorescence was low. In time-lapse experiments, CFP and YFP bleached at different rates. The ratio was corrected for bleaching using a method described elsewhere²¹. Because ruffles and cell edges can move rapidly, it was important to exclude motion artefacts; we found the same result when the order of acquisition of CFP and FRET images was reversed or when fixed cells were analysed (data not shown).

Received 4 October 2005; accepted 21 February 2006.

Published online 19 March 2006.

1. Raftopoulos, M. & Hall, A. Cell migration: Rho GTPases lead the way. *Dev. Biol.* **265**, 23–32 (2004).
2. Burridge, K. & Wennerberg, K. Rho and Rac take center stage. *Cell* **116**, 167–179 (2004).
3. Fukata, Y. *et al.* Phosphorylation of adducin by Rho-kinase plays a crucial role in cell motility. *J. Cell Biol.* **145**, 347–361 (1999).
4. Palazzo, A. F., Cook, T. A., Alberts, A. S. & Gundersen, G. G. mDia mediates Rho-regulated formation and orientation of stable microtubules. *Nature Cell Biol.* **3**, 723–729 (2001).
5. Nimmual, A. S., Taylor, L. J. & Bar-Sagi, D. Redox-dependent downregulation of Rho by Rac. *Nature Cell Biol.* **5**, 236–241 (2003).
6. Sander, E. E., ten Klooster, J. P., van Delft, S., van der Kammen, R. A. & Collard, J. G. Rac downregulates Rho activity: reciprocal balance between both GTPases determines cellular morphology and migratory behaviour. *J. Cell Biol.* **147**, 1009–1022 (1999).
7. Ren, X. D., Kiosses, W. B. & Schwartz, M. A. Regulation of the small GTP-binding protein Rho by cell adhesion and the cytoskeleton. *EMBO J.* **18**, 578–585 (1999).
8. Griesbeck, O., Baird, G. S., Campbell, R. E., Zacharias, D. A. & Tsien, R. Y. Reducing the environmental sensitivity of yellow fluorescent protein. Mechanism and applications. *J. Biol. Chem.* **276**, 29188–29194 (2001).
9. Miyawaki, A. & Tsien, R. Y. Monitoring protein conformations and interactions by fluorescence resonance energy transfer between mutants of green fluorescent protein. *Methods Enzymol.* **327**, 472–500 (2000).
10. Hoffman, G. R., Nassar, N. & Cerione, R. A. Structure of the Rho family GTP-binding protein Cdc42 in complex with the multifunctional regulator RhoGDI. *Cell* **100**, 345–356 (2000).
11. Miura, K. *et al.* ARAP1: a point of convergence for Arf and Rho signaling. *Mol. Cell* **9**, 109–119 (2002).
12. Norman, J. C. *et al.* ARF1 mediates paxillin recruitment to focal adhesions and potentiates Rho-stimulated stress fiber formation in intact and permeabilized Swiss 3T3 fibroblasts. *J. Cell Biol.* **143**, 1981–1995 (1998).
13. Cho, S. Y. & Klemke, R. L. Purification of pseudopodia from polarized cells reveals redistribution and activation of Rac through assembly of a CAS/Crk scaffold. *J. Cell Biol.* **156**, 725–736 (2002).
14. Matsumura, F., Ono, S., Yamakita, Y., Totsukawa, G. & Yamashiro, S. Specific localization of serine 19 phosphorylated myosin II during cell locomotion and mitosis of cultured cells. *J. Cell Biol.* **140**, 119–129 (1998).
15. Worthylake, R. A., Lemoine, S., Watson, J. M. & Burridge, K. RhoA is required for monocyte tail retraction during transendothelial migration. *J. Cell Biol.* **154**, 147–160 (2001).
16. Kurokawa, K. & Matsuda, M. Localized RhoA activation as a requirement for the induction of membrane ruffling. *Mol. Biol. Cell* **16**, 4294–4303 (2005).
17. Shaw, R. J., Henry, M., Solomon, F. & Jacks, T. RhoA-dependent phosphorylation and relocalization of ERM proteins into apical membrane/actin protrusions in fibroblasts. *Mol. Biol. Cell* **9**, 403–419 (1998).
18. Fernandez-Borja, M., Janssen, L., Verwoerd, D., Hordijk, P. & Neeffjes, J. RhoB regulates endosome transport by promoting actin assembly on endosomal membranes through Dia1. *J. Cell Sci.* **118**, 2661–2670 (2005).
19. Merrifield, C. J. *et al.* Endocytic vesicles move at the tips of actin tails in cultured mast cells. *Nature Cell Biol.* **1**, 72–74 (1999).
20. Ridley, A. J., Paterson, H. F., Johnston, C. L., Diekmann, D. & Hall, A. The small GTP-binding protein rac regulates growth factor-induced membrane ruffling. *Cell* **70**, 401–410 (1992).
21. Hodgson, L., Nalbant, P., Shen, F. & Hahn, K. M. Imaging and photobleach correction of Mero-CBD, sensor for endogenous Cdc42 activation. *Methods Enzymol.* **406**, 140–156 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to G. Bokoch, P. Sun, M. Schwartz, R. Tsien, C. Der and E. Sahai for reagents, and to F. Shen for help with image analysis. This work was supported by grants from the Swiss National Science Foundation, Roche Research Foundation, Novartis and Philip Morris to O.P., and from the National Institutes of Health to K.M.H. and R.L.K.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to K.M.H. (khahn@med.unc.edu) or O.P. (pertz@ucsd.edu).

Pharmaco-metabonomic phenotyping and personalized drug treatment

T. Andrew Clayton¹, John C. Lindon¹, Olivier Cloarec¹, Henrik Antti², Claude Charuel³, Gilles Hanton³, Jean-Pierre Provost³, Jean-Loïc Le Net³, David Baker⁴, Rosalind J. Walley⁵, Jeremy R. Everett⁵ & Jeremy K. Nicholson¹

There is a clear case for drug treatments to be selected according to the characteristics of an individual patient, in order to improve efficacy and reduce the number and severity of adverse drug reactions^{1,2}. However, such personalization of drug treatments requires the ability to predict how different individuals will respond to a particular drug/dose combination. After initial optimism, there is increasing recognition of the limitations of the pharmacogenomic approach, which does not take account of important environmental influences on drug absorption, distribution, metabolism and excretion^{3–5}. For instance, a major factor underlying inter-individual variation in drug effects is variation in metabolic phenotype, which is influenced not only by genotype but also by environmental factors such as nutritional status, the gut microbiota, age, disease and the co- or pre-administration of other drugs^{6,7}. Thus, although genetic variation is clearly important, it seems unlikely that personalized drug therapy will be enabled for a wide range of major diseases using genomic knowledge alone. Here we describe an alternative and conceptually new ‘pharmaco-metabonomic’ approach to personalizing drug treatment, which uses a combination of pre-dose metabolite profiling and chemometrics to model and predict the responses of individual subjects. We provide proof-of-principle for this new approach, which is sensitive to both genetic and environmental influences, with a study of paracetamol (acetaminophen) administered to rats. We show pre-dose prediction of an aspect of the urinary drug metabolite profile and an association between pre-dose urinary composition and the extent of liver damage sustained after paracetamol administration.

¹H nuclear magnetic resonance (NMR) spectroscopy has been widely applied as a metabolite profiling tool for metabonomic studies, as it enables many endogenous metabolites to be quantified rapidly and reproducibly without derivatization or separation^{8–11}. In one of many potential applications, NMR-based metabonomic analysis of post-dose rodent biofluids has been developed as a rapid and non-invasive means of assessing the toxicity of potential drug compounds, and this has involved studying the effects of a variety of model toxins¹². In one such study, we administered galactosamine hydrochloride (800 mg kg⁻¹) to a group of ten rats and found the extent of the induced liver effects to be so variable that the rats could be classified as either ‘responders’ or ‘non-responders’. Searching for the cause of this variation, we performed principal component analysis (PCA) on the NMR spectra of the relevant pre-dose urine samples and observed some discrimination between responder and non-responder groups in terms of their pre-dose metabolite profiles (Fig. 1a). Although this result was not sufficient to say that galactosamine responder/non-responder behaviour is

predictable, it suggested that information on individual responses to xenobiotics might be contained in the metabolite patterns of pre-dose biofluids. We thus conceived the possibility of ‘pharmaco-metabonomics’, which we define as ‘the prediction of the outcome (for example, efficacy or toxicity) of a drug or xenobiotic intervention in an individual based on a mathematical model of pre-intervention metabolite signatures’.

We also conducted a larger study on the severity of liver damage induced in rats by allyl alcohol (50 mg kg⁻¹), and found a weak but statistically significant association between the extent of the induced damage and the pre-dose urinary data (Fig. 1b), although in this case the discriminating factor was the total pre-dose excretion of organic compounds as estimated by ¹H NMR spectroscopy, rather than the detailed metabolite pattern. However, knowing that the multivariate endogenous metabolite profiles of biofluids reflect inter-subject variation with respect to a multitude of factors such as age, sex,

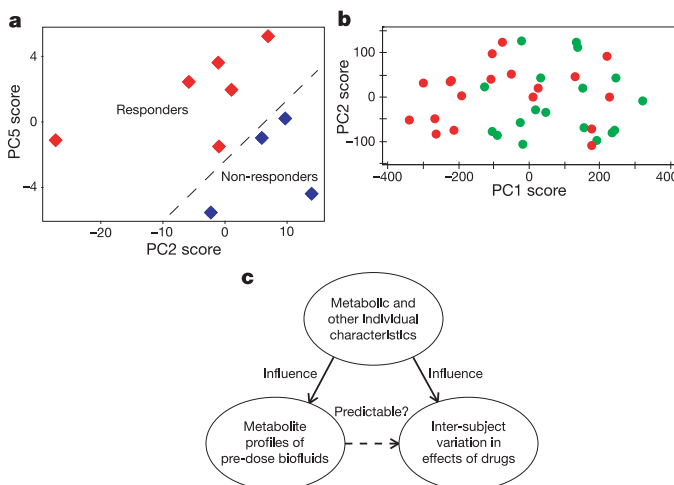


Figure 1 | Preliminary studies and the pharmaco-metabonomic hypothesis.

a, A scores plot from PCA of the ¹H NMR spectra of urine samples obtained before dosing male Sprague–Dawley rats with galactosamine hydrochloride (800 mg kg⁻¹). Each point represents an individual rat, with colour-coding according to post-dose behaviour. **b**, A scores plot from PCA of urinary data obtained before dosing male Sprague–Dawley rats with allyl alcohol (50 mg kg⁻¹). Each point represents an individual rat, with colour-coding according to the extent of post-dose liver damage. Red signifies greatest damage, green signifies least damage, mid-damage group excluded. **c**, Schematic of our pharmaco-metabonomic hypothesis.

¹Biological Chemistry, Biomedical Sciences Division, Faculty of Natural Sciences, Sir Alexander Fleming Building, Imperial College London, South Kensington, London SW7 2AZ, UK. ²Department of Chemistry, Umeå University, 901 87 Umeå, Sweden. ³Pfizer Global Research & Development, Centre de Recherche, 37401 Amboise Cedex, France. ⁴Pfizer Inc., 2800 Plymouth Road, Ann Arbor, Michigan 48105, USA. ⁵Pfizer Global Research & Development, Ramsgate Road, Sandwich, Kent CT13 9NJ, UK.

diet, ethnicity and disease^{13–17}, and recognizing that many of the factors that influence the metabolism of drugs would normally operate on endogenous and dietary substances, we decided to test the hypothesis that the pre-dose metabolite profile of an individual animal contains sufficient information to allow the prediction of aspects of drug metabolism and toxicity in that animal without any prior knowledge of the animal's genomic profile (Fig. 1c). We chose the commonly used analgesic paracetamol (acetaminophen) for this investigation.

We collected pre- and post-dose urine samples from 65 rats given a single toxic-threshold dose of paracetamol (600 mg kg⁻¹, a treatment that resulted in no mortality or clinical signs), and analysed all of these samples by ¹H NMR spectroscopy (Fig. 2). The post-dose spectra showed characteristic signals of both endogenous and paracetamol-related metabolites, with the major paracetamol-related metabolites being identified as paracetamol sulphate (S), paracetamol glucuronide (G), the mercapturic acid (MA) derived from paracetamol, and paracetamol (P) itself. Using these spectra in conjunction with the known urinary volumes, we determined the amounts and relative proportions of paracetamol-related metabolites excreted by each animal, and modelled the variation in these data in relation to the multivariate endogenous metabolite profiles obtained from the pre-dose samples. Liver damage of variable severity was identified and quantified by clinical chemistry and histopathology in samples taken at approximately 24 h post-dose, with the extent of the microscopically visible damage being scored in each of five liver lobes and a mean histology score (MHS) derived for each rat (see Supplementary Fig. 1 and Supplementary Tables). The inter-animal variation in MHS was also modelled in relation to the multivariate endogenous metabolite profiles obtained from the pre-dose samples, in an attempt to integrate classical end-point histopathology with a pre-dose metabonomic data set. We then tested the predictive ability of the various models.

The mole ratio of paracetamol glucuronide to paracetamol (G/P) was found to be the most convincingly predicted of the various post-dose metabolite quantities. Thus, we built and validated a PLS model (projection to latent structure; see Methods) for predicting, from their pre-dose metabolite profiles, the G/P values obtained post-dose for individual animals (Fig. 3). The most important factors underlying this model were a positive correlation ($r = 0.48$) between G/P and the integral of the $\delta 5.06$ – 5.14 region of the pre-dose NMR

spectra, and negative correlations ($r = -0.56$ and -0.54) between G/P and the integrals of the $\delta 8.98$ – 9.10 and $\delta 0.50$ – 0.86 regions of the pre-dose spectra (Fig. 3b). The integrals for the latter two regions were correlated to one another ($r = 0.90$), but showed no clear relationship to the integral for the $\delta 5.06$ – 5.14 region. Inspection of the pre-dose spectra revealed some small but distinct signals in the $\delta 5.06$ – 5.14 region, with the $\delta 8.98$ – 9.10 region and much of the $\delta 0.50$ – 0.86 region being relatively featureless but positive relative to the zero line.

Our validation (Fig. 3c, d) confirms that the predictive model for G/P is robust, and the positive correlation between G/P and the $\delta 5.06$ – 5.14 region of the pre-dose spectra is consistent with metabolic control. The glucuronide part of paracetamol glucuronide produces a doublet in this region post-dose, and the signals in the $\delta 5.06$ – 5.14 region of the pre-dose spectra may arise, correspondingly, from endogenous ether glucuronides. It is logical that G/P would be positively correlated to the pre-existing tendency of each individual rat to form and excrete such ether glucuronides, and the integral for the $\delta 5.06$ – 5.14 region could be a good reflection of that tendency, being relatively free of other signals. The significance of the negative correlations between G/P and the specified regions of the pre-dose spectra is not understood, but might reflect some dependency on urinary protein.

Although we did not obtain a fully validated model for predicting post-dose histology, our analysis did show a statistically significant association between the nature of the pre-dose urinary biochemical profile and the post-dose histological outcome. Having assigned each animal to one of three histology classes (described in Table 1), PCA was carried out on the pre-dose spectral data (62 animals), with partial separation between histology classes 1 and 3 being found on principal component 2 (PC2) (Fig. 4a). Furthermore, a weak but statistically significant correlation ($r = -0.34$; $P = 0.007$) was found between the PC2 scores and MHS (Fig. 4b). PCA was also carried out

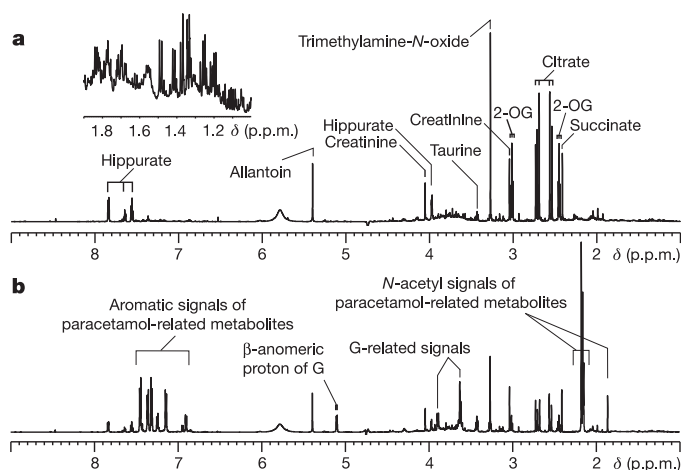


Figure 2 | Representative ¹H NMR spectra. **a**, ¹H NMR spectra of pre-dose (**a**, -48 h to -24 h) and post-dose (**b**, 0 to $+24$ h) urine samples from a rat dosed with paracetamol (600 mg kg⁻¹). The inset in **a** is an expansion of the $\delta 1.9$ – 1.0 region, indicating the complexity of the endogenous profile and the richness of the embedded information. 2-OG, 2-oxoglutarate; G, paracetamol glucuronide.

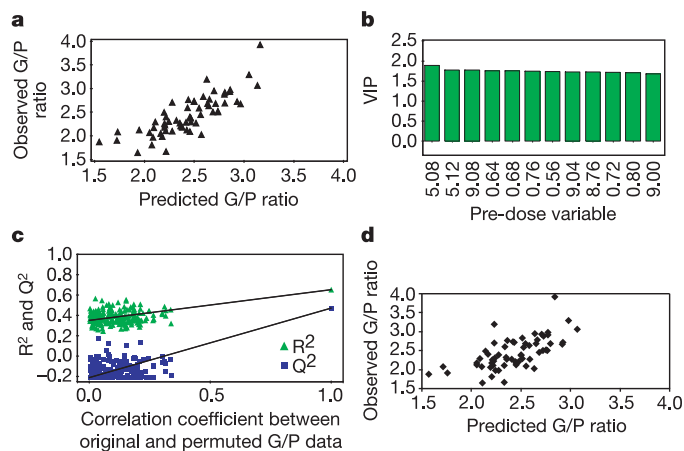


Figure 3 | Pre-dose prediction of the urinary mole ratio of paracetamol glucuronide to paracetamol (G/P) obtained in paracetamol-dosed rats. **a**, Observed versus predicted G/P values for a two-component PLS model in which all predictions relate to model-building data. **b**, The twelve regions of the pre-dose ¹H NMR spectra most important to the above model. Each of the 0.04 p.p.m.-wide spectral segments is identified by the chemical shift at its mid-point. VIP, variable influence on projection. **c**, Internal validation of the above model, showing clear decreases in performance as the G/P data are permuted relative to the pre-dose data. R^2 describes how well the derived model fits the data, and is the proportion of the sum of squares explained by the model. Q^2 describes the predictive ability of the derived model and is the cross-validated R^2 . **d**, The result of a seven-round cross-validation exercise in which every point represents test data not used in the model-building (regression line is represented by the equation: observed value = $(0.89 \times \text{predicted value}) + 0.29$; root mean square error of predictions 0.32).

on the pre-dose data for classes 1 and 3 only (32 animals), with partial separation of those classes again being observed on PC2 (Fig. 4c) and the statistical significance of that separation being confirmed by a Mann–Whitney *U*-test ($P = 0.002$).

The main pre-dose factors underlying the histology class discrimination are the levels of taurine, trimethylamine-*N*-oxide (TMAO) and betaine (Fig. 4d). A higher pre-dose level of taurine is associated more with class 1 than with class 3, whereas a higher combined pre-dose level of TMAO and betaine is associated more with class 3 than with class 1. The beneficial relationship found between pre-dose urinary taurine and the extent of paracetamol-induced liver damage is consistent with earlier findings for the degree of liver damage induced by a variety of other liver toxins, and with the protective effect observed when taurine is administered to rats before or soon after a hepatotoxic dose of paracetamol^{18,19}. The significance of the pre-dose level of taurine might lie in its known defensive properties²⁰, but urinary taurine might also reflect the availability of inorganic sulphate (to which taurine is metabolically related), which is a precursor of the paracetamol-sulphating agent phosphoadenosine phosphosulfate (PAPS)^{21,22}. The latter interpretation is consistent with our observation that most of the animals with a higher degree of liver necrosis (MHS > 2.5) showed a low proportion of paracetamol sulphate in their post-dose urine. In contrast, a higher pre-dose level of TMAO is associated with more paracetamol-induced liver damage, suggesting that the gut bacteria might have a role in determining the extent of such damage^{23,24}—and the contribution of the gut bacteria to a bayesian probabilistic ('Pachinko') model of metabolism has already been postulated¹¹. The occasional pre-dose presence of a significant quantity of betaine, which produces an NMR signal that overlaps that of TMAO, may be a confounding factor and certainly contributed to the abnormal position of one of the class 1 animals on the scores plot shown in Fig. 4c.

Whatever the basis of the observed pre-dose discrimination, our paracetamol study findings identify statistically significant relation-

ships between variation in the pre-dose data and the post-dose variation in histopathology and in the urinary level of a drug metabolite relative to its parent. Thus, the two apparently independent models provide the first demonstration of the concept of pharmaco-metabonomics, in which drug-induced responses in individuals are potentially predictable from their pre-dose metabolite profiles, which serve as biochemical signatures that report simultaneously on multiple factors of relevance to drug metabolism and drug effects.

Looking forward, we propose that this pharmaco-metabonomic approach, which amounts to response-targeted pre-dose phenotyping, might provide the basis of a future population-screening tool for selecting individuals according to their suitability for treatment with particular drugs, drug classes or drug doses. Furthermore, this approach should have certain advantages over methods for drug and dose selection that are dependent on the use of test compounds²⁵ to characterize particular aspects of an individual's metabolic phenotype. In particular, the pharmaco-metabonomic approach has the potential to provide automatic identification and weighting of multiple response-determining factors. A further inherent benefit of the pharmaco-metabonomic approach is the identification of new biomarkers that are predictive of individual responses.

The main potential application we envisage for pharmaco-metabonomics is with respect to personalized human healthcare, and there is clearly a need for further development and to investigate how well the approach can be transferred from single-strain laboratory animals to human subjects, for whom much greater genetic and environmental variation would be expected. Recent developments in data analysis should assist pharmaco-metabonomic modelling and biomarker identification^{26,27}. Other analytical techniques such as LC–MS and GC–MS might also be used, with the likelihood of detecting a large number of low-concentration metabolites not normally accessible by NMR. Metabolite profiling of fluids other than urine, such as blood and faecal extracts, should also provide additional information. In practice, the success or otherwise of the pharmaco-metabonomic approach would be expected to vary from drug to drug, and would depend on the nature of the challenge posed by each drug, on the response of interest and on the extent to which the relevant response-controlling factors are reported in the pre-dose data. However, in principle, by using this methodology, adverse drug reactions could potentially be avoided and drugs and dose levels could be targeted more effectively according to the metabolic and other characteristics of each individual. We envisage that optimal personalization of drug treatments may eventually involve a variety of response-prediction approaches, including both pharmaco-metabonomics and pharmacogenomics. However, pharmaco-metabonomics has an important theoretical advantage over pharmacogenomics in that it can potentially take account of both genomic and environmental factors affecting drug-induced responses. Furthermore, although pharmaco-metabonomics would normally relate to predicting drug- or xenobiotic-induced responses, we envisage that similar methodology could also be applied

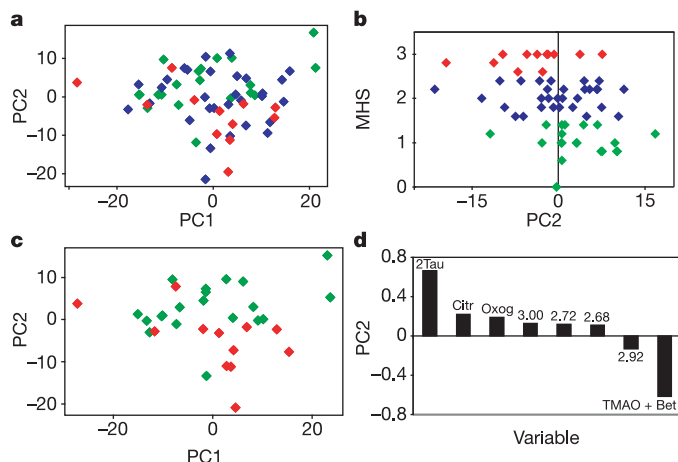


Figure 4 | Pre-dose discrimination of the degree of liver damage obtained in paracetamol-dosed rats. **a**, A scores plot from PCA of the pre-dose NMR data. Each point represents a single rat and is colour-coded by its histology class (with increasing severity of damage, class 1 is green, class 2 is blue, class 3 is red; see Table 1). **b**, Plot of mean histology score (MHS) versus the PC2 score obtained from the above PCA, with colour-coding as before. **c**, A scores plot from PCA of the pre-dose NMR data for rats in histology classes 1 and 3. Each point represents a single rat, with colour-coding as before. **d**, A loadings plot corresponding to **c**, showing the variables making the largest contributions to PC2, and the direction of each contribution. Individual 0.04 p.p.m.-wide spectral segments are identified by the chemical shifts at their midpoints, and variables corresponding to particular compounds are identified by name. Tau, taurine; Citr, citrate; Oxog, 2-oxoglutarate; TMAO, trimethylamine-*N*-oxide; Bet, betaine. '2Tau' indicates doubling of the Tau values.

Table 1 | Liver histopathology in paracetamol-dosed rats at about 24 h after dosing

Mean histology score (MHS)*	Extent of liver necrosis	Number of rats per category	Class designation and colour-coding for PCA
MHS = 0	No significant necrosis	1	1, green
0.0 < MHS < 1.5	Minimal necrosis	20	1, green
1.5 < MHS < 2.5	Mild necrosis	32	2, blue
2.5 < MHS	Moderate necrosis	12	3, red

* It was not possible to obtain mean histology scores of 1.5 or 2.5 because of the methodology used, in which individual scores across five liver lobes were averaged. See Supplementary Information for details.

to predicting individual responses to broader medical, dietary, microbiological or physiological challenges.

METHODS

This section relates to the paracetamol study only. Further details are provided in the Supplementary Methods.

Animal treatment and sampling. The paracetamol study was performed in accordance with relevant national legislation using 75 male Sprague-Dawley rats (CrI:CD (SD)IGS BR) obtained from Charles River. The rats (approximately seven weeks old) were placed in individual cages in a controlled environment, with free access to water and a commercial feed. Sixty-five animals received, by oesophageal intubation, a single oral dose of paracetamol (600 mg kg^{-1}) as an aqueous suspension containing methylcellulose (0.5% w/v) and Tween 80 (0.1% w/v). A further ten rats were used as a control group and were orally dosed with vehicle only. Individual pre-dose (−48 to −24 h) and post-dose (0–24 h) urine samples were collected into ice-cooled vessels containing 0.5 ml of an aqueous 100 mg ml^{-1} solution of sodium azide as a preservative. After post-dose urine collection, individual blood samples were collected and centrifuged to recover plasma for clinical chemistry. The rats were then killed using CO_2 and the liver of each animal was examined, weighed and sampled for histopathology.

NMR sample preparation, spectral acquisition, processing and construction of data sets. Urine samples were prepared as described in the Supplementary Information. ^1H NMR spectra were acquired (7200 Hz spectral width, 65,536 time domain points, 8 dummy scans, 64 real scans) at 600 MHz, at a nominal 303 K, on a Bruker DRX 600 NMR spectrometer operated by XWINNMR software (both Bruker Biospin) with the 'noesyprsat' pulse sequence²⁸ used to suppress the water signal during a 3-s relaxation delay and during the 0.1-s mixing time. After Fourier transformation with 0.3-Hz line broadening and a single zero-filling, pre-dose spectra were manually phased and baseline-corrected, and the chemical shift scale set by assigning the value of δ 0 to the signal from the added TSP (sodium 3-trimethylsilyl-[2,2,3,3- $^2\text{H}_4$]-1-propionate). Each of these processed spectra was then 'data-reduced' using AMIX software (Bruker Biospin), where the spectral regions $\delta > 9.5$, $\delta 6.1$ – 5.5 , $\delta 5.0$ – 4.5 and $\delta < 0.5$ were discarded before dividing the remainder of each spectrum into sequential segments ('bins') of 0.04 p.p.m. width and obtaining an integral for each segment. These integrals were then normalized to give the same total integral for each data-reduced spectrum. Each bin was initially identified by the chemical shift at its midpoint, but quantities relating to certain compounds were derived by making appropriate bin combinations as described in the Supplementary Information.

Post-dose spectra were processed as above and subsequently with resolution enhancement, in order to determine the amounts and the relative proportions of the paracetamol-related compounds^{29,30} excreted by each animal, by reference to the cluster of *N*-acetyl signals in the range $\delta 2.11$ – 2.22 , the TSP signal and the post-dose urinary volume. The amounts excreted were corrected to unit body mass.

Clinical chemistry. Blood plasma samples were analysed at 30 °C on an AU600 clinical analyser (Olympus) for a variety of parameters, and univariate statistical analyses were performed. Details and results are provided in the Supplementary Information.

Histopathology. For each animal, ten representative samples of the liver (two each from the left, right, left middle, right middle and caudate lobes) were examined, the changes in each lobe scored, a mean histology score (MHS) calculated and a histology class assigned as described in Table 1.

Multivariate modelling. Pirouette (v.2.7 and 3.1, Infometrix) and SIMCA P+ (v.10.0 and 10.5, Umetrics) software packages were used. After initial investigations, three animals were excluded from all subsequent modelling because of abnormal pre- or post-dose behaviour.

Predictive multivariate modelling was carried out in SIMCA, with a supervised pattern-recognition method (projection to latent structure, PLS) being used to model the co-variation between NMR-derived pre-dose data (*X*) and selected post-dose response variables (*Y*). Model validation was performed as described in the Supplementary Information. VIP (variable influence on projection) values characterize the relative overall importance of the individual *X* variables to the model, and the weights for each component of the model reveal the nature of the relationship between each *X* variable and the predicted quantity, *Y*.

For predictive modelling of the G/P mole ratio, total area-normalized pre-dose spectral data were used with unit variance variable scaling. An unsupervised pattern recognition method (principal component analysis, PCA) was also used. Histology-coded PCA was performed, using Pirouette, on the pre-dose spectral data normalized to constant total spectral area, with mean-centred variable scaling and with the histology class of each animal assigned as described in

Table 1. The relevant loadings describe how the original variables contribute to each principal component.

Univariate statistical analysis and tests of correlation. All significance tests were two-sided. See Supplementary Information for details.

Received 17 October 2005; accepted 14 February 2006.

1. Spear, B. B., Heath-Chiozzi, M. & Huff, J. Clinical application of pharmacogenetics. *Trends Mol. Med.* **7**, 201–204 (2001).
2. Pagliarulo, V., Datar, R. H. & Cole, R. J. Role of genetic and expression profiling in pharmacogenomics: the changing face of patient management. *Curr. Issues Mol. Biol.* **4**, 101–110 (2002).
3. Ingelman-Sundberg, M. Genetic variability in susceptibility and response to toxicants. *Toxicol. Lett.* **120**, 259–268 (2001).
4. Nebert, D. W., Jorge-Nebert, L. & Vesell, E. S. Pharmacogenomics and "individualized drug therapy": high expectations and disappointing achievements. *Am. J. Pharmacogenomics* **3**, 361–370 (2003).
5. van Aken, J., Schmedders, M., Feuerstein, G. & Kollek, R. Prospects and limits of pharmacogenetics: the thiopurine methyl transferase (TPMT) experience. *Am. J. Pharmacogenomics* **3**, 149–155 (2003).
6. Rang, H. P., Dale, M. M. & Ritter, J. M. *Pharmacology* (Churchill Livingstone, Edinburgh, 1995).
7. Tannock, G. W. *Normal Microflora: An Introduction to Microbes Inhabiting the Human Body* (Chapman and Hall, London, 1995).
8. Nicholson, J. K., Lindon, J. C. & Holmes, E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR data. *Xenobiotica* **29**, 1181–1189 (1999).
9. Lindon, J. C., Nicholson, J. K., Holmes, E. & Everett, J. R. Metabonomics: metabolic processes studied by NMR spectroscopy of biofluids combined with pattern recognition. *Concepts Magn. Reson.* **12**, 289–320 (2000).
10. Nicholson, J. K., Connelly, J., Lindon, J. C. & Holmes, E. Metabonomics: a platform for studying drug toxicity and gene function. *Nature Rev. Drug Discov.* **1**, 153–161 (2002).
11. Nicholson, J. K. & Wilson, I. D. Understanding 'global' systems biology: metabonomics and the continuum of metabolism. *Nature Rev. Drug Discov.* **2**, 668–676 (2003).
12. Lindon, J. C. et al. The Consortium for Metabonomic Toxicology (COMET): aims, activities and achievements. *Pharmacogenomics* **6**, 691–699 (2005).
13. Shockcor, J. P. & Holmes, E. Metabonomic applications in toxicity screening and disease diagnosis. *Curr. Top. Med. Chem.* **2**, 35–51 (2002).
14. Rawson, E. S., Clarkson, P. M., Price, T. B. & Miles, M. P. Differential response of muscle phosphocreatine to creatine supplementation in young and old subjects. *Acta Physiol. Scand.* **174**, 57–65 (2002).
15. Zuppi, C. et al. ^1H NMR spectra of normal urines: reference ranges of the major metabolites. *Clin. Chim. Acta* **265**, 85–97 (1997).
16. Zoratti, R. A review on ethnic differences in plasma triglycerides and high-density-lipoprotein cholesterol: is the lipid pattern the key factor for the low coronary heart disease rate in people of African origin? *Eur. J. Epidemiol.* **14**, 9–21 (1998).
17. Lenz, E. M. et al. Metabonomics, dietary influences and cultural differences: a ^1H NMR-based study of urine samples obtained from healthy British and Swedish subjects. *J. Pharm. Biomed. Anal.* **36**, 841–849 (2004).
18. Waterfield, C. J., Turton, J. A., Scales, M. D. C. & Timbrell, J. A. The correlation between urinary and liver taurine levels and between pre-dose urinary taurine and liver damage. *Toxicology* **77**, 1–5 (1993).
19. Waters, E. et al. Role of taurine in preventing acetaminophen-induced hepatic injury in the rat. *Am. J. Physiol. Gastrointest. Liver Physiol.* **280**, G1274–G1279 (2001).
20. Timbrell, J. A., Seabra, V. & Waterfield, C. J. The *in vivo* and *in vitro* protective properties of taurine. *Gen. Pharmacol.* **26**, 453–462 (1995).
21. Coomes, M. W. in *Textbook of Biochemistry with Clinical Correlations* (ed. Devlin, T. M.) 473–474 (Wiley-Liss, New York, 1997).
22. Liu, L. & Klaassen, C. D. Different mechanisms of saturation of acetaminophen sulfate conjugation in mice and rats. *Toxicol. Appl. Pharmacol.* **139**, 128–134 (1996).
23. Smith, J. L., Wishnok, J. S. & Deen, W. M. Metabolism and excretion of methylamines in rats. *Toxicol. Appl. Pharmacol.* **125**, 296–308 (1994).
24. Nicholls, A. W., Mortishire-Smith, R. J. & Nicholson, J. K. NMR spectroscopic-based metabonomic studies of urinary metabolite variation in acclimatizing germ-free rats. *Chem. Res. Toxicol.* **16**, 1395–1404 (2003).
25. Grant, D. M., Tang, B. K. & Kalow, W. A simple test for acetylator phenotype using caffeine. *Br. J. Clin. Pharmacol.* **58**, S788–S793 (2004).
26. Cloarec, O. et al. Evaluation of the O-PLS limitations caused by chemical shift variability and improved visualization of biomarker changes in ^1H NMR spectroscopic metabonomic studies. *Anal. Chem.* **77**, 517–526 (2005).
27. Cloarec, O. et al. Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic ^1H NMR data sets. *Anal. Chem.* **77**, 1282–1289 (2005).
28. Claridge, T. D. W. (ed.) *High-Resolution NMR Techniques in Organic Chemistry. Tetrahedron Organic Chemistry Series 19* (Elsevier, Amsterdam, 1999).

29. Bales, J. R., Sadler, P. J., Nicholson, J. K. & Timbrell, J. A. Urinary excretion of acetaminophen and its metabolites as studied by proton NMR spectroscopy. *Clin. Chem.* **30**, 1631–1636 (1984).
30. Gregus, Z., Madhu, C. & Klaassen, C. D. Species variation in toxication and detoxication of acetaminophen in vivo: a comparative study of biliary and urinary excretion of acetaminophen metabolites. *J. Pharmacol. Exp. Ther.* **244**, 91–99 (1988).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the technical staff of Pfizer Global R&D in Amboise, France, for their assistance in performing the animal-related work, and thank R. L. Smith for discussions. T.A.C. thanks Pfizer for financial support. O.C. acknowledges the support of the Wellcome Trust (BAIR project).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.K.N (j.nicholson@imperial.ac.uk).

LETTERS

Designed divergent evolution of enzyme function

Yasuo Yoshikuni^{1,4}, Thomas E. Ferrin^{1,5} & Jay D. Keasling^{1,2,3,4}

It is generally believed that proteins with promiscuous functions divergently evolved to acquire higher specificity and activity^{1–5}, and that this process was highly dependent on the ability of proteins to alter their functions with a small number of amino acid substitutions (plasticity)⁶. The application of this theory of divergent molecular evolution to promiscuous enzymes may allow us to design enzymes with more specificity and higher activity. Many structural and biochemical analyses have identified the active or binding site residues important for functional plasticity (plasticity residues)^{6–10}. To understand how these residues contribute to molecular evolution, and thereby formulate a design methodology, plasticity residues were probed in the active site of the promiscuous sesquiterpene synthase γ -humulene synthase^{11,12}. Identified plasticity residues were systematically recombined based on a mathematical model in order to construct novel terpene synthases, each catalysing the synthesis of one or a few very different sesquiterpenes. Here we present the construction of seven specific and active synthases that use different reaction pathways to produce the specific and very different products. Creation of these enzymes demonstrates the feasibility of exploiting the underlying evolvability of this scaffold, and provides evidence that rational approaches based on these ideas are useful for enzyme design.

Promiscuous enzyme activities have long been believed to be an important determinant for molecular evolution of more specific and active enzyme functions^{1–5}. It is thought that primordial enzymes may have been promiscuous to render primitive organisms adaptable to their environment; enzymes with higher specificity are the result of divergent evolution (gene duplications and subsequent mutations), driven by selective pressure, from promiscuous precursor enzymes. Numerous biochemical analyses have suggested that proteins have an ability to improvise novel or altered functions with a small number of amino acid substitutions (plasticity)^{5–10}. We refer to those residues that primarily govern enzyme specificity as plasticity residues. The underlying evolvability of promiscuous enzymes is also thought to be very important in molecular evolution in order to allow organisms to adapt rapidly in response to environmental changes⁶. Because natural evolution is known to be a highly accomplished designer for protein function, understanding how proteins acquire novel or altered functions and how plasticity residues contribute to the natural evolution process may help to formulate an efficient design methodology for new enzymes.

To investigate how promiscuous proteins might evolve to acquire more active and specific functions, we chose as a model enzyme γ -humulene synthase, a sesquiterpene synthase from *Abies grandis*^{11,12} that is known to produce 52 different sesquiterpenes from a sole substrate, farnesyl diphosphate, through a wide variety of cyclization mechanisms (Fig. 1 and Supplementary Fig. 1). All terpene synthases share a similar active site scaffold^{13–16}. The reaction is initiated by cleavage of the diphosphate group to yield a carbocation intermediate, which is then cyclized into many different

structures. In general, this reaction generates a large number of terpene structures with different regio- and stereochemistries¹⁷. In none of the previous work was it shown that one could successfully control this reaction pathway of a highly reactive carbocation species, or improve the product selectivity for chemically more complex reactions.

Owing to the extreme promiscuity of γ -humulene synthase, the product distribution could be very sensitive to changes in specific amino acid residues; hence, the enzyme should be an excellent model to study plasticity residues. However, the significantly greater promiscuity of this enzyme compared to other enzymes makes functional design more challenging, and the lack of a selection or a high-throughput screen for evolved terpene synthases makes directed evolution nearly impossible. In addition, it is extremely difficult to predict the relationships between primary sequence and enzyme function in this class of enzymes, because enzymes are closely related within or near species regardless of their functional disparity^{18–20}. This lack of relatedness among the limited number of known sesquiterpene synthases with a similar function makes functional design based on phylogenetic analysis nearly impossible²¹. Therefore, a method that would allow one to predict the effect of changes in amino acid residues in terpene synthases on product selectivity would alleviate the need for a high-throughput screen or genomic analysis in designing synthases useful for mass production of single terpenes that have found use as drugs, flavours, fragrances, nutraceuticals and in many other applications^{17,22}.

Although plasticity residues can be found anywhere in proteins, many biochemical and genomic analyses have indicated that they tend to be more focused inside or near active sites^{6–10}. Consistent with this observation, results from directed evolution²³, which can search for such residues over entire proteins, also indicate that these residues most often occur in the active site^{6,9}. To determine the active-site residues important for γ -humulene synthase function, a homology structure for γ -humulene synthase was first built using the crystal structure of 5-*epi*-aristolochene synthase (Protein Data Bank entry 5eat¹⁴) as a guide^{24,25}. Although mutations to residues in the conserved aspartate-rich motif in the active site are known to alter the reaction mechanisms of terpene synthases^{12,26}, these residues were not considered further because mutations in this motif are usually accompanied by significant losses of activity^{12,26}. As a result, the 19 residues composing the active-site contour were selected for saturation mutagenesis to investigate how each residue contributes to a particular reaction mechanism (Fig. 2). Saturation mutagenesis of residue S484 and subsequent screening by gas chromatography-mass spectrometry (GC-MS) suggested that 80 mutants were sufficient to obtain almost all possible amino acid changes (Supplementary Table 1). The altered product distribution from each mutant was normalized to that of the wild-type enzyme and profiled. Although many of these residues were identified to be plastic, four residues significantly affected catalysis: W315, M447, S484 and Y566 (Supplementary Figs 2–5). Mutations to these residues shifted the relative

¹UCSF/UCB Joint Graduate Group in Bioengineering, ²Department of Chemical Engineering, and ³California Institute for Quantitative Biomedical Research (QB3), University of California at Berkeley, Berkeley, California 94720, USA. ⁴Synthetic Biology Department, Physical Bioscience Division, Lawrence Berkeley National Laboratory, Berkeley, California 94710, USA. ⁵Department of Pharmaceutical Chemistry and Biopharmaceutical Sciences, University of California at San Francisco, San Francisco, California 94143, USA.

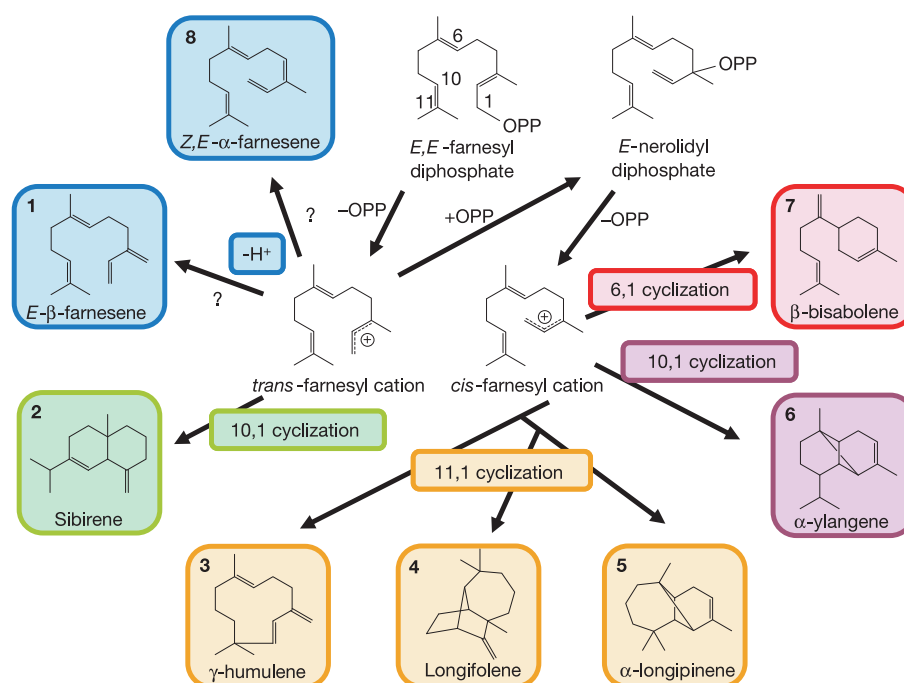


Figure 1 | γ -Humulene synthase cyclization reaction mechanisms. When the substrate, farnesyl diphosphate, binds to the enzyme active site via divalent magnesium cations, the diphosphate group is released to yield either *trans*- or *cis*-farnesyl cation. From the *trans*-farnesyl cation, sibirane (2) is produced by the 10,1 cyclization reaction. From the *cis*-farnesyl cation, γ -humulene (3), longifolene (4) and α -longipinene (5) are produced

through an 11,1 cyclization reaction; α -ylangene (6) through a 10,1 cyclization reaction; and β -bisabolene (7) through a 6,1 cyclization reaction. *E*- β -farnesene (1) and *Z,E*- α -farnesene (8) can be produced by directed deprotonation from either farnesyl cation (see Supplementary Fig. 1 for more details).

selectivity (the amount of one product relative to another product) by 100- to 1,000-fold.

To investigate further how these plasticity residues contribute to molecular evolution and to formulate a design methodology for product selectivity, the mutations were systematically recombined based on the profiles obtained from saturation mutagenesis. The recombination was carried out using an algorithm (detailed in the Methods) based on the assumption that each plasticity residue is independent—the effect of a particular mutation on the reaction mechanism should be the same for the wild-type enzyme and any mutants. Thus, a set of mutations to achieve a desired enzyme function was predicted based on how much the product distribution moved towards the desired product distribution, as measured by ω for a particular combination of mutations. For example, in the

construction of a β -bisabolene synthase (BBA; product 7 in Fig. 1), two mutants (M447H/A336V/I562T ($\omega = 6.7$) and M447H/A336V/I562V ($\omega = 5.4$)) were predicted to be the best combinations of mutations to these three amino acids to maximize β -bisabolene selectivity ($\omega = 24.4$ for wild type). M447 was important in specifying the 10,1 or 6,1 closure from the *trans*- or *cis*-farnesyl cation ($\omega = 16.3$, Fig. 3b, c); A336 was important in specifying the 11,1 closure from the *cis*-farnesyl cation ($\omega = 10.8$, Fig. 3d, e); and I562 was important in specifying acyclic terpene formation (Fig. 3f). M447H/A336V/I562T reduced production of product 1 and showed 2.5-fold better selectivity for production of 7 over 1 compared to that of M447H/A336V/I562V (Fig. 3f; see also Supplementary Table 2). Thus, a β -bisabolene synthase was successfully constructed while maintaining its activity (Table 1).

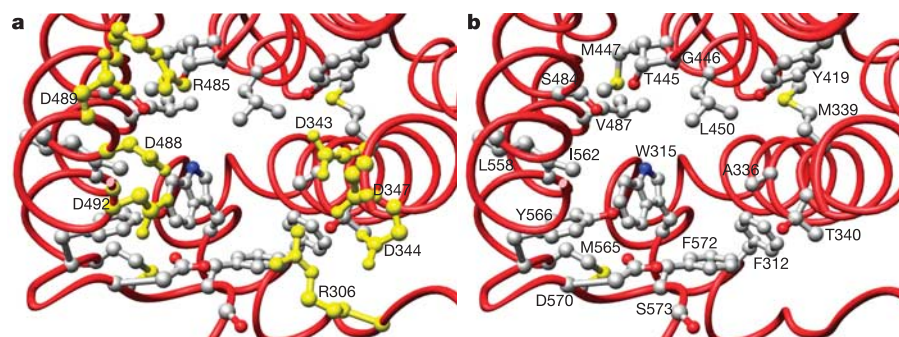


Figure 2 | The homology structural model for the γ -humulene synthase active site. **a**, The residues that were not considered are shown in yellow. Six aspartate residues in two different aspartate-rich motifs and two arginine residues, which are generally conserved in all sesquiterpene cyclases, were not considered, because these residues are thought to be catalytically important, and mutations to these residues would have decreased enzyme

activity significantly. **b**, The 19 residues in the active site are shown, and these residues were targeted for saturation mutagenesis and systematic remodelling. (See Supplementary Fig. 11 for the primary sequence alignment and Supplementary Data 1 for the three-dimensional coordinates for the homology model of γ -humulene synthase.)

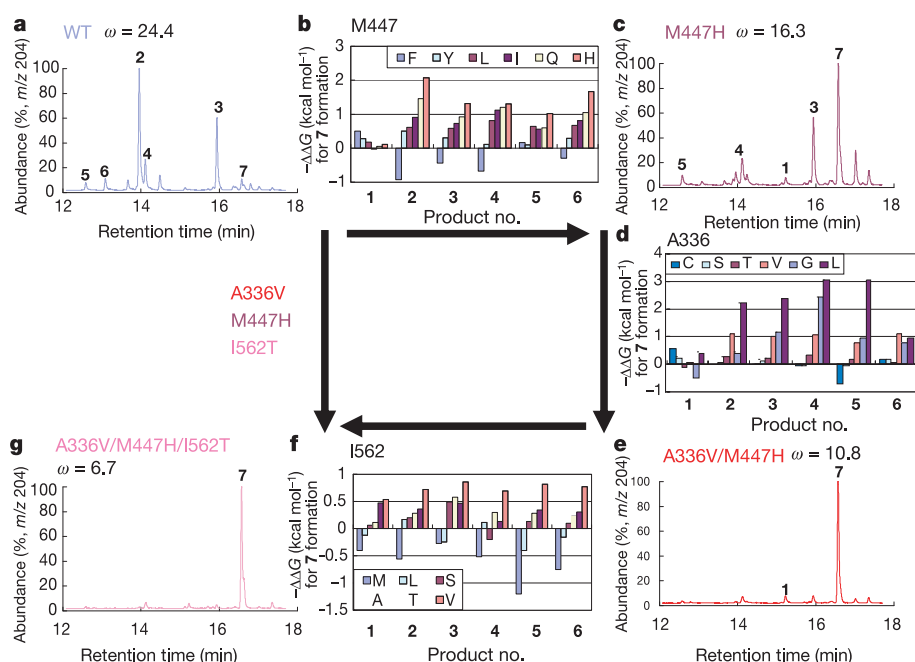


Figure 3 | Systematic remodelling of plasticity residues to design β -bisabolene synthase. Chromatograms in **a**, **c**, **e** and **g** show the GC-MS analysis of terpene production for wild type, M447H, A336V/M447H and A336V/M447H/I562T, respectively. The numbers for each peak correspond to those in Fig. 1. **b**, **d** and **f** show the free energy change for formation of 7

over that for formation of 1–6 for each successive change to residues M447, A336 and I562 compared to that of wild type, respectively. Quantitative analyses are shown in Supplementary Table 2. Mutations were sequentially added, so as to reduce ω . β -Bisabolene synthase was successfully constructed from γ -humulene synthase through A336V/M447H/I562T.

We used this same method to create an *E*- β -farnesene (1)/*Z*,*E*- α -farnesene (8) synthase (BFN: W315P), a sibirane (2) synthase (SIB: F312Q/M339A/M447F), a longifolene (4) synthase (LFN: A336S/A336V/I562V), an α -longipinene (5) synthase (ALP: A336C/T445C/S484C/I562L/M565L) and two new γ -humulene (3) synthases (HUM: M339N/S484C/M565I and AYG: S484A/Y566F), the latter having significantly improved α -ylangene (6) production (Fig. 4, Table 1, Supplementary Figs 6–10 and Supplementary Tables 3–7). SIB, HUM, ALP and AYG are new synthases that have not yet been discovered in nature. Although the construction of α -ylangene synthase was not achieved (AYG), the algorithm predicted that it would not be possible to create such an enzyme from the current set of plasticity residues.

Almost all mutations were added by saturation mutagenesis so as not to miss any better substitutions that were not predicted by the algorithm. However, in almost all cases, the predicted substitutions

gave the desired product distribution (Fig. 3; see also Supplementary Tables 2–7 and Supplementary Figs 6–10). In general, all of the designed enzymes maintained a level of specific activity comparable to the wild-type ancestor (Table 1). These results indicate that in order to generate a specific enzyme from another specific enzyme, the enzyme must first acquire promiscuous function, supporting the theory that specific enzymes could evolve from promiscuous precursor enzymes with surprisingly few mutations⁶. In addition, we observed convergent evolution, as the product profiles of HUM and AYG are very similar to each other despite the significant differences in mutations that gave rise to these specific functions.

Although we assumed that the plasticity residues behaved as if they were independent and that changes to these residues were additive, the effects could be partially additive, synergistic, antagonistic, or even absent altogether²⁷. If two residues to be mutated do not interact, then the effect of mutating the residues is likely to be

Table 1 | Summary for wild-type γ -humulene synthase and its derivatives

Clones*	Mutations	Product distributions (%)†								Yield (times) ‡	k_{cat} (s ⁻¹)	K_m (μ M)	k_{cat}/K_m (M ⁻¹ s ⁻¹)
		1	2	3	4	5	6	7	8				
Wild type	Wild type	3.0	23.1	45.1	13.4	4.7	3.8	6.9	ND	1	$2.36 \pm 0.16 (\times 10^{-2})$	4.66 ± 0.97	5.07×10^3
BFN	W315P	54.1	2.9	2.2	ND	ND	ND	6.7	34.0	2.1	$1.94 \pm 0.04 (\times 10^{-3})$	0.179 ± 0.024	1.08×10^4
SIB	F312Q, M339A, M447F	6.4	78.1	11.4	2.6	0.1	1.0	0.4	ND	1.8	$4.63 \pm 0.19 (\times 10^{-4})$	3.01 ± 0.54	1.54×10^2
HUM	M339N, S484C, M565I	2.7	0.4	85.7	3.4	3.5	4.3	ND	ND	1.2	$1.81 \pm 0.09 (\times 10^{-3})$	2.08 ± 0.53	8.70×10^2
LFN	A317NS, A336S, S484C, I562V	1.7	1.4	12.6	63.0	12.6	3.2	5.5	ND	4.4	$6.96 \pm 0.50 (\times 10^{-2})$	3.83 ± 0.95	1.81×10^4
ALP	A336C, T445C, S484C, I562L, M565L	7.7	ND	11.7	13.3	61.5	2.2	3.6	ND	13	$3.81 \pm 0.18 (\times 10^{-3})$	4.59 ± 0.84	8.31×10^2
AYG	S484A, Y566F	2.6	0.4	54.6	3.5	15.8	14.7	8.4	ND	3.3	$1.21 \pm 0.06 (\times 10^{-2})$	6.05 ± 1.04	1.99×10^3
BBA	A336V, M447H, I562T	6.4	0.4	3.8	4.7	1.1	0.1	83.6	ND	4.2	$2.24 \pm 0.10 (\times 10^{-2})$	2.88 ± 0.40	7.77×10^3

BFN, *E*- β -farnesene synthase; SIB, sibirane synthase; HUM, γ -humulene synthase; LFN, longifolene synthase; ALP, α -longipinene synthase; AYG, α -ylangene over-producer (another γ -humulene synthase); BBA, β -bisabolene synthase; ND, production not detected; K_m , Michaelis constant.

* All constructs are made based on a soluble variant (data not shown).

† All product distributions were represented for 1–8 as 100%; these products generally correspond to more than 85–95% and to 75% of total products in mutants and wild type, respectively. All product distributions were determined from triplicates, and standard deviations were lower than 2%.

‡ *In vivo* productivity over wild type for each desired product.

§ A317N occurred during recombination, and improved *in vivo* terpene production without a change in product distribution.

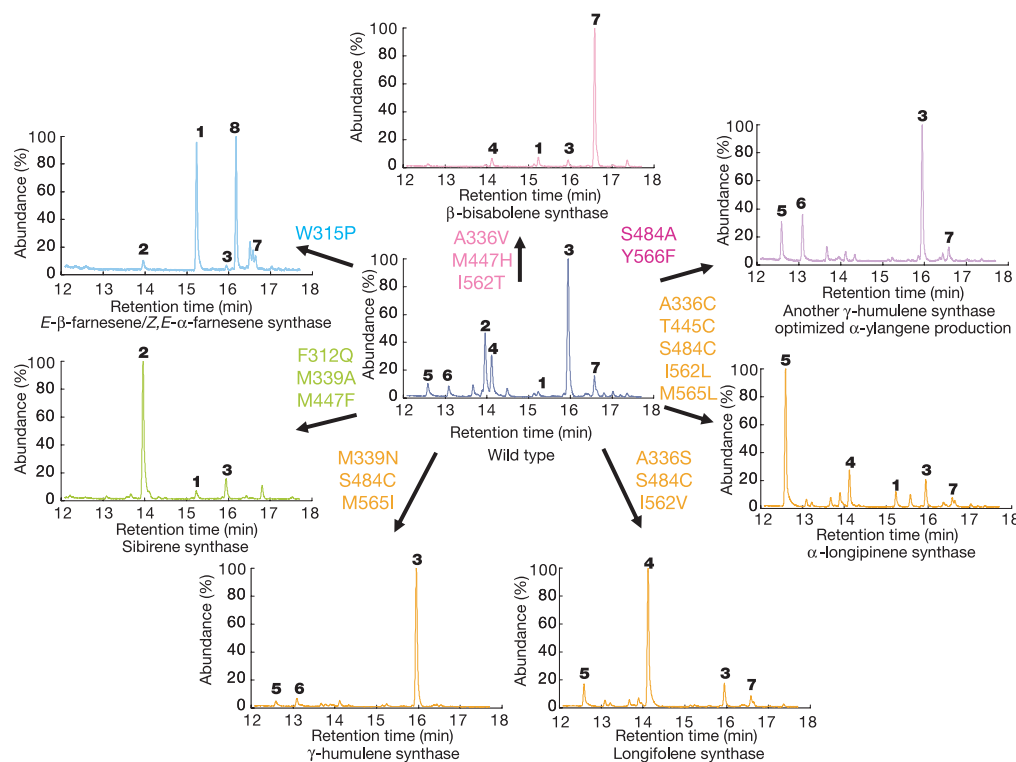


Figure 4 | Divergent evolution of novel sesquiterpene synthases from γ -humulene synthase. Chromatograms show the GC-MS analysis for sesquiterpene production from both wild type (centre) and variants of γ -humulene synthase. The numbers for each peak and the colours for each chromatogram correspond to those in Fig. 1. All γ -humulene synthase variants were designed based on the systematic remodelling and constructed

by site-directed saturation mutagenesis and site-directed mutagenesis. Primary sequence and reaction mechanism relationships were clearly observed. For example, enzymes that produce **3**, **4** and **5**, all of which are produced from the 11,1 cyclization pathway, are more closely related (sharing S484C) to each other than to any of the other enzymes.

additive. As direct or indirect interactions between two residues increase, the impact of multiple mutations may be far from additive²⁷. All effects other than antagonistic, however, would still be predicted to some extent using the methodology outlined here. Interestingly, with the exception of the double mutant M339N/S484C, all mutations introduced into the enzyme were effectively additive. Hence, in practice, the assumption that mutations will have an additive effect is rational, and the resulting enzyme design methodology is simple, yet powerful.

Construction of the large number of sesquiterpene synthases required fewer than 2,500 mutants to be screened. However, in directed evolution—currently touted as an effective tool to alter protein function^{6,9,10,23,28}—tens of thousands to a million or more mutants must be screened to find a few critical mutations; hence its application is limited by the availability of an efficient screening method. The systematic recombination approach described herein enabled us to design enzyme specificity rapidly and efficiently without a screen for the desired activity. Because plasticity residues have very important roles in proteins²⁹, this approach, with some modifications, may be useful for designing novel functions for many other proteins, including enzymes, protein ligands/receptors, transcription factors and antibodies.

On the basis of the theories of molecular evolution and experimental observations, we formulated an approach for systematic recombination of the promiscuous γ -humulene synthase. We successfully constructed a large number of novel specific sesquiterpene synthases, each producing one or a few products derived from a predominant reaction pathway while largely maintaining the specific activity of the original enzyme. These results suggest that: (1) divergent evolution by rational design may be feasible on a significantly larger scale than currently possible; (2) plasticity residues could significantly drive molecular evolution; and (3) most of the

substitutions in plasticity residues additively affect protein functions. Although we demonstrated systematic recombination using the subset of plasticity residues located in the active site, other residues can also be considered in order to construct other specific enzymes or even enzymes that produce unnatural products.

METHODS

GC-FID and GC-MS analysis for sesquiterpenes. A single colony harbouring pTrcHUM and pBBRMBIS (kanamycin; antibiotic-resistance gene was replaced)²² was inoculated into Luria Bertani (LB) medium containing 50 $\mu\text{g ml}^{-1}$ carbenicillin and 50 $\mu\text{g ml}^{-1}$ kanamycin and grown overnight at 30 °C. An aliquot (50 μl) of this seed culture was inoculated into fresh LB medium containing 10 mM mevalonate, 0.1 mM isopropyl-1-thio- β -D-galactopyranoside (IPTG), 50 $\mu\text{g ml}^{-1}$ carbenicillin and 50 $\mu\text{g ml}^{-1}$ kanamycin (5 ml), overlaid with 500 μl dodecane, and grown for 24 h at 30 °C. To screen the library resulting from site-directed saturation mutagenesis, a single colony harbouring only pTrcHUM was inoculated into LB medium containing 0.1 mM IPTG and 50 $\mu\text{g ml}^{-1}$ carbenicillin overlaid with 500 μl dodecane, and grown for 24 h at 30 °C. An aliquot of dodecane (50 μl) was diluted into 200 μl of ethyl acetate, and the mixture was analysed by GC-MS using a GC oven temperature programme of 80 °C for 1 min, then steps of 30 °C min^{-1} to 110 °C, 5 °C min^{-1} to 160 °C and 130 °C min^{-1} to 250 °C for Cyclosil-B capillary column analysis, or at 80 °C for 3 min, then steps of 5 °C min^{-1} to 160 °C and 120 °C min^{-1} to 300 °C for DB-5MS capillary column analysis. Gas chromatography-flame ionization detector (GC-FID) analysis was also carried out to quantify each sesquiterpene product using the method described herein. The proportion of each product was determined based on the ratio of the relative peak abundance for each product. Sesquiterpenes were identified from their mass spectra and GC retention times by comparison to available authentic standards (γ -humulene, longifolene, α -longipinene and *E*- β -farnesene) and spectra in libraries previously reported in the literature (sibirene¹¹, α -ylangene³⁰, β -bisabolene³⁰ and *Z,E*- α -farnesene³⁰).

Homology structural modelling of γ -humulene synthase. The homology structural model for γ -humulene synthase (Supplementary Data 1) was built

using MODELLER²⁵ (<http://salilab.org/modeller/>). The alignment (Supplementary Fig. 11) and the structure of 5-*epi*-aristolochene synthase¹⁴ (PDB entry 5eat) were used as guides. The resulting homology structure was visualized using Chimera (<http://www.cgl.ucsf.edu/chimera>).

Algorithm for systematic remodelling of plasticity residues. To design the specificity for novel sesquiterpene synthases, combinations of mutations were selected based on the results from the previous screening. Assuming that there is no interaction between plasticity residues, the effect of a certain mutation is the same for both wild type and other mutants. Therefore, the product distribution profile upon another round of mutagenesis can easily be calculated using the following equation:

$$D_i = \frac{d_i x_i}{\sum_{j=1}^n d_j x_j} \times 100 (\%)$$

where D_i is the predicted percentage of product distribution of compound i for all compounds 1 to n (in this study $n = 17$), d_i is percentage of parent product distribution of compound i (which can be predicted) and x_i is the effect of a particular mutation on compound i productivity relative to the wild-type enzyme.

The predicted percentage of product distribution of compound i by addition of m distinctive mutations (m th generation), $D_{i,m}$, is then represented as follows:

$$D_{i,m} = \frac{d_{i,0} \prod_{g=1}^m x_{i,g}}{\sum_{j=1}^n d_{j,0} \prod_{g=1}^m x_{j,g}} \times 100 (\%)$$

where $d_{i,0}$ represents the parent (0th generation) product distribution of compound i (which can be predicted) and $x_{i,m}$ represents the effects by the m th mutation. To select the mutations that probably introduce the desired function, the root mean square deviation of the predicted product distribution from the desired product distribution was calculated using the following equation:

$$\omega_m = \sqrt{\frac{\sum_{i=1}^n (D_{i,m} - d'_i)^2}{n}} (\%)$$

where ω_m represents the relative closeness of product profiles of mutants with m th generation (m mutations) to the one specified, and d'_i is the percentage of desired product distribution for compound i (for example, for β -bisabolene synthase $d'_7 = 100\%$ and $d'_{i \neq 7} = 0\%$). To improve the accuracy of the design methodology, we included 17 different product profiles, which correspond to 95% of total products.

To maintain specific activity and productivity, the overall productivity was calculated using the following equation:

$$P = \sum_{i=1}^n P_i = \sum_{i=1}^n p_i x_i$$

where P is total productivity, P_i is predicted productivity for compound i , and p_i is parent productivity for compound i . Similarly, the overall productivity at the m th generation was calculated using the following equation:

$$P_m = \sum_{i=1}^n P_{i,m} = \sum_{i=1}^n p_i \prod_{g=1}^m x_{i,g}$$

The combinations of mutations were selected so as to decrease the ω_m value and to maintain P_m . The selected mutations were added to γ -humulene synthase and the results were inspected. See Supplementary Methods for additional methods.

Received 25 October 2005; accepted 26 January 2006.

Published online 22 February; corrected 27 February 2006.

- Jensen, R. A. Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**, 409–425 (1976).
- O'Brien, P. J. & Herschlag, D. Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol.* **6**, R91–R105 (1999).
- Copley, S. D. Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Curr. Opin. Chem. Biol.* **7**, 265–272 (2003).
- James, L. C. & Tawfik, D. S. Conformational diversity and protein evolution—a 60-year-old hypothesis revisited. *Trends Biochem. Sci.* **28**, 361–368 (2003).
- Gerlt, J. A., Babbitt, P. C. & Rayment, I. Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity. *Arch. Biochem. Biophys.* **433**, 59–70 (2005).
- Aharoni, A. *et al.* The 'evolvability' of promiscuous protein functions. *Nature Genet.* **37**, 73–76 (2005).
- Bone, R., Silen, J. L. & Agard, D. A. Structural plasticity broadens the specificity of an engineered protease. *Nature* **339**, 191–195 (1989).
- van Den Heuvel, R. H., Fraaije, M. W., Ferrer, M., Mattevi, A. & van Berkel, W. J.

Inversion of stereospecificity of vanillyl-alcohol oxidase. *Proc. Natl Acad. Sci. USA* **97**, 9455–9460 (2000).

- Matsumura, I. & Ellington, A. D. *In vitro* evolution of β -glucuronidase into a β -galactosidase proceeds through non-specific intermediates. *J. Mol. Biol.* **305**, 331–339 (2001).
- Aharoni, A. *et al.* Directed evolution of mammalian paraoxonases PON1 and PON3 for bacterial expression and catalytic specialization. *Proc. Natl Acad. Sci. USA* **101**, 482–487 (2004).
- Steele, C. L., Crock, J., Bohlmann, J. & Croteau, R. Sesquiterpene synthases from grand fir (*Abies grandis*). Comparison of constitutive and wound-induced activities, and cDNA isolation, characterization, and bacterial expression of δ -selinene synthase and γ -humulene synthase. *J. Biol. Chem.* **273**, 2078–2089 (1998).
- Little, D. B. & Croteau, R. B. Alteration of product formation by directed mutagenesis and truncation of the multiple-product sesquiterpene synthases δ -selinene synthase and γ -humulene synthase. *Arch. Biochem. Biophys.* **402**, 120–135 (2002).
- Lesburg, C. A., Zhai, G., Cane, D. E. & Christianson, D. W. Crystal structure of pentalene synthase: mechanistic insights on terpenoid cyclization reactions in biology. *Science* **277**, 1820–1824 (1997).
- Starks, C. M., Back, K., Chappell, J. & Noel, J. P. Structural basis for cyclic terpene biosynthesis by tobacco 5-*epi*-aristolochene synthase. *Science* **277**, 1815–1820 (1997).
- Caruthers, J. M., Kang, I., Rynkiewicz, M. J., Cane, D. E. & Christianson, D. W. Crystal structure determination of aristolochene synthase from the blue cheese mold, *Penicillium roqueforti*. *J. Biol. Chem.* **275**, 25533–25539 (2000).
- Rynkiewicz, M. J., Cane, D. E. & Christianson, D. W. Structure of trichodiene synthase from *Fusarium sporotrichioides* provides mechanistic inferences on the terpene cyclization cascade. *Proc. Natl Acad. Sci. USA* **98**, 13543–13548 (2001).
- Glasby, J. S. *Encyclopedia of the Terpenoids* (Wiley, Chichester/New York, 1982).
- Aubourg, S., Lecharny, A. & Bohlmann, J. Genomic analysis of the terpenoid synthase (AtTPS) gene family of *Arabidopsis thaliana*. *Mol. Genet. Genomics* **267**, 730–745 (2002).
- Martin, D. M., Faldt, J. & Bohlmann, J. Functional characterization of nine Norway spruce TPS genes and evolution of gymnosperm terpene synthases of the TPS-d subfamily. *Plant Physiol.* **135**, 1908–1927 (2004).
- Bohlmann, J., Meyer-Gauen, G. & Croteau, R. Plant terpenoid synthases: molecular biology and phylogenetic analysis. *Proc. Natl Acad. Sci. USA* **95**, 4126–4133 (1998).
- Lichtarge, O., Bourne, H. R. & Cohen, F. E. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358 (1996).
- Martin, V. J., Pitera, D. J., Withers, S. T., Newman, J. D. & Keasling, J. D. Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. *Nature Biotechnol.* **21**, 796–802 (2003).
- Stemmer, W. P. Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* **370**, 389–391 (1994).
- Sali, A. Comparative protein modeling by satisfaction of spatial restraints. *Mol. Med. Today* **1**, 270–277 (1995).
- Baker, D. & Sali, A. Protein structure prediction and structural genomics. *Science* **294**, 93–96 (2001).
- Rynkiewicz, M. J., Cane, D. E. & Christianson, D. W. X-ray crystal structures of D100E trichodiene synthase and its pyrophosphate complex reveal the basis for terpene product diversity. *Biochemistry* **41**, 1732–1741 (2002).
- Mildvan, A. S. Inverse thinking about double mutants of enzymes. *Biochemistry* **43**, 14517–14520 (2004).
- Joo, H., Lin, Z. & Arnold, F. H. Laboratory evolution of peroxide-mediated cytochrome P450 hydroxylation. *Nature* **399**, 670–673 (1999).
- James, L. C., Roversi, P. & Tawfik, D. S. Antibody multispecificity mediated by conformational diversity. *Science* **299**, 1362–1367 (2003).
- Adams, R. P. *Identification of Essential Oil Components by Gas Chromatography/Mass Spectroscopy* (Allured Publishing Corporation, Carol Stream, Illinois, 1995).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We would like to thank P. C. Babbitt, J. D. Newman, M. C. Chang and S. C.-H. Pegg for discussions and critical reading of the manuscript. We are also grateful for D. Herschlag for critical comments. This research was funded by the Bill & Melinda Gates Foundation, the US Department of Agriculture, and the National Science Foundation.

Author Contributions Y.Y. and J.D.K. conceived the project; Y.Y., J.D.K. and T.E.F. designed the experiments; and Y.Y. and J.D.K. wrote the paper.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.D.K. (keasling@berkeley.edu).

-  FOCUS
-  SPOTLIGHT
-  RECRUITMENT
-  ANNOUNCEMENTS
-  EVENTS

naturejobs

Physical exertion

The 1900s are largely viewed as the century of physics — and, by the same token, many believe we are now living in the century of biology. If that is true, then someone has forgotten to tell US graduate students. The number of graduates pursuing physics is swelling, and their career prospects are looking bright, according to a report issued by the American Institute of Physics last month.

The report confirms that there has been a shift within the field. For the period 1997–2001, more foreign graduate students than US nationals enrolled in physics programmes. But since 1998, the number of US citizens enrolling has gone up by 47%, and in 2004, US nationals made up 55% of first-year graduate enrolments in physics and astronomy.

For foreign physics students enrolling in the United States, there has also been a shift in their country of origin. Although China has always led the pack in terms of numbers — it has now increased that lead: in 2004 one-third of foreign graduate students enrolling in physics were Chinese. This gain has largely been at the expense of Europe where the proportion of students enrolling fell from 36% in 1999 to 21% in 2004.

What is driving the revival of physics among US students? It is hard to say for sure. The number of first degrees being awarded in the subject has risen in recent years, which is married to a perception that physics offers a viable career path.

Sadly, some of that optimism is likely to be tempered in time, when aspirations meet reality. More than half of all the graduate students will be aiming for positions at research universities but there simply aren't that many positions available — even if the much-rumoured mass retirement of ageing faculty members comes to pass. As a result, some of this year's class may well find themselves working as stock-market analysts, information-technology professionals — or even side-by-side with biologists, in one of the increasingly interdisciplinary life-science labs.



Paul Smaglik, *Naturejobs* editor

CONTACTS

Publisher: Ben Crowe
Editor: Paul Smaglik
Assistant Editor: Corie Lok

European Head Office, London
 The Macmillan Building,
 4 Crinan Street
 London N1 9XW, UK
 Tel: +44 (0) 20 7843 4961
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

European Sales Manager:
 Andy Douglas (4975)
Natureevents: Silke Opstrup (4994)

UK/RoW/Ireland/Italy:
 Nils Moeller (4953)
 Irene Viglia-Atton (4944)
Scandinavia/Spain/Portugal:
 Evelina Rubio Håkansson (4973)
France/Switzerland/Belgium:
 Amelie Pequignot (4974)
**Germany/Austria/
 The Netherlands:**
 Reya Silao (4970)

Advertising Production Manager:
 Billie Franklin
 To send materials use London
 address above.
 Tel: +44 (0) 20 7843 4814

Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com
Naturejobs web development: Tom Hancock
Naturejobs online production: Niamh Shields

**European Satellite Office
 Germany/Austria/
 The Netherlands:**
 Patrick Phelan
 Tel: +49 89 54 90 57 11
 Fax: +49 89 54 90 57 20
 e-mail: p.phelan@nature.com

US Head Office, New York
 75 Varick Street,
 9th Floor, New York,

NY 10013-1917
 Tel: +1 800 989 7718
 Fax: +1 800 989 7103
 e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo
 Chiyoda Building,
 2-37 Ichigayatamachi,
 Shinjuku-ku,
 Tokyo 162-0843
 Tel: +81 3 3267 8751
 Fax: +81 3 3267 8746
Asia-Pacific Sales Director: Rinoko Asami
 e-mail: rasami@naturejpn.com

Seeing the big picture

Creating a new drug is a long and painstaking process, involving the skills and talents of numerous types of scientist, says **Hannah Hoag**. Each is vital to different stages of producing a drug that's both safe and effective.

It's a long and painstaking process, steering a drug through the preclinical phase of development. A promising drug candidate has plenty of potential when it leaves the discovery laboratory, but that's just the first step, and the statistics are against it. Only five of the 5,000 compounds that enter preclinical testing go on to human trials. And it takes a team with many different skills to create the drug, check for safety and scale it up for production.

During preclinical development, scientists carry out the process chemistry and scale-up, perform thorough toxicology studies, and sometimes decide on the drug's final formulation. In this phase, scientists contribute their technical and creative talents to move a promising candidate from the research lab into the clinic. In preclinical development, companies weed out drugs that are too cumbersome to produce, or that may be harmful, saving millions of dollars on wasted clinical trials.

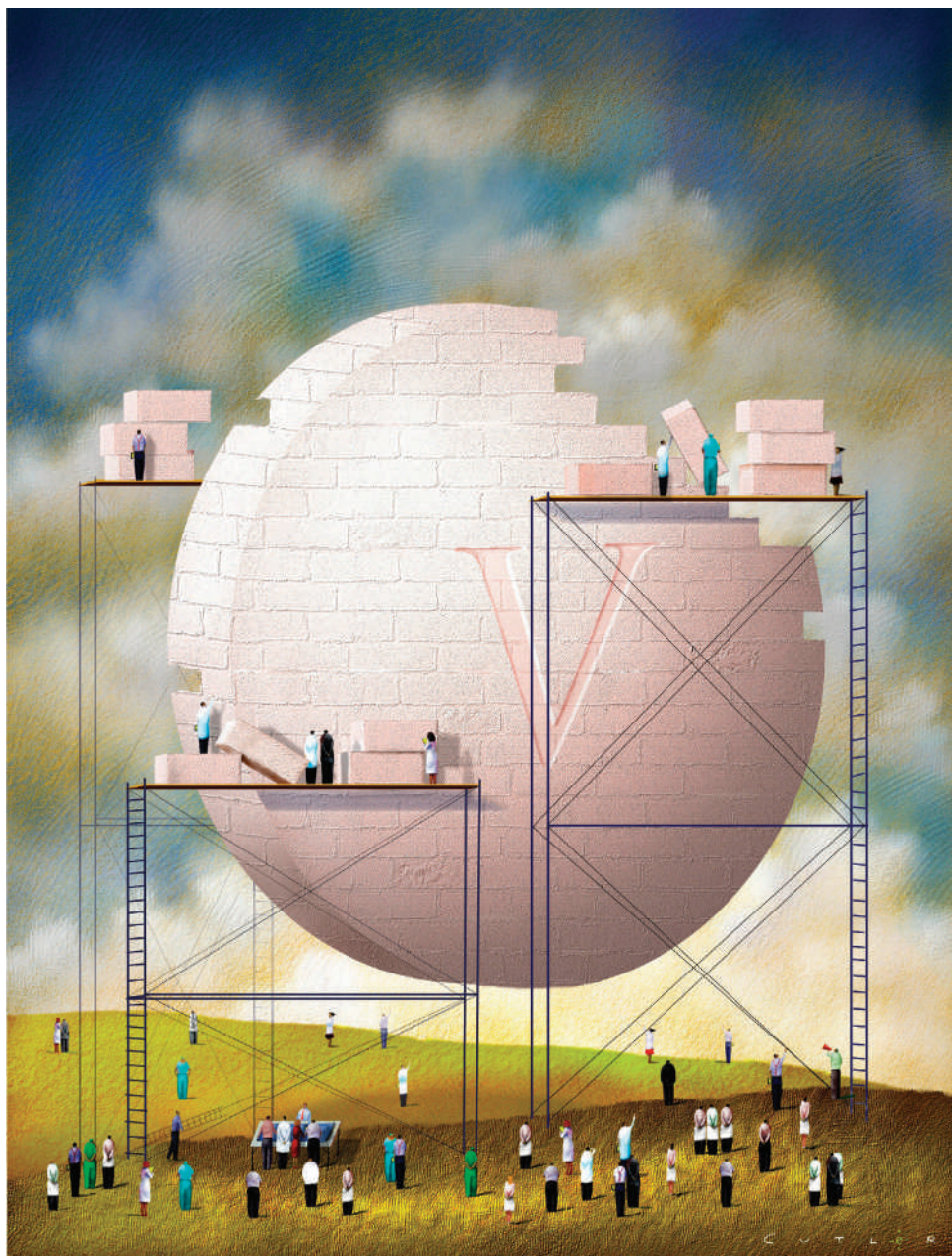
Right place, right time

Preclinical development is a hot place to be. Organic growth, mergers, acquisitions and in-licensing are adding early-stage drug candidates to companies' pipelines. According to the American Chemical Society's annual survey, chemical-industry employment in the pharmaceutical sector grew slightly between 2000 and 2005. Although growth has stagnated since early 2004, job loss due to consolidation may be turning. Companies that shed staff after mergers are now hiring again, thanks to growth in their pipelines and movement towards later-stage compounds. And unlike manufacturing jobs, US preclinical development positions have yet to be outsourced abroad.

Moving a drug from discovery to preclinical stage means increasing its production from milligrams to kilograms. Doing so requires finding efficient ways to scale up. "In medicinal chemistry, you're simply interested in getting hold of the compound and you're not always bothered at this stage of the R&D process by how expensive the chemistry is," says



The right mix: small firms need a collection of skills.



David Nicholson, executive vice-president of research and development at Organon International in Roseland, New Jersey.

Scaling up production uses the talents of several types of chemist. Synthetic chemists devise scale-up processes, aiming to keep costs low. Analytical chemists characterize the new compound. And chemical engineers — most in demand at biotech companies — find the best ways to produce enzymes. All three types play critical roles on the road towards clinical trials.

The synthetic chemist asks critical questions about production, says Ken Batchelor, senior vice-president of GlaxoSmithKline's Center of Excellence for Drug Discovery: Metabolic and Viral Diseases. "Can the substance be produced in fewer steps, more quickly or with a greater yield?" Even when times are tight, there is always demand for highly skilled synthetic chemists.

Analytical chemists assess the compound and identify any impurities. Characterizing the product and assuring its quality are such important steps that there are sometimes as many analytical chemists and

CORBIS

biochemists as there are process chemists, says Chuck Goochee, senior director of bioprocess research and development at Merck.

This phase also emphasizes the safety of chemical reactions, says Scott Biller, head of global discovery chemistry at the Novartis Institutes for BioMedical Research. Heat-producing, or exothermic, reactions, barely noticeable in discovery, can blossom into serious problems during development, where skilled process chemists with industry experience can trouble-shoot. And concerns over waste and environmental costs are prompting industry leaders to seek cleaner, more efficient ways to produce drugs (see *Nature* **440**, 378–379; 2006).

Biological-based therapeutics such as proteins and monoclonal antibodies pose their own challenges in scaling up. In research settings, scientists use cell expression systems to produce small amounts of therapeutic molecules. But these methods have such low productivity that they're not commercially viable, so chemical engineers search for the best system to produce larger amounts.

Then they work with molecular biologists, biochemists and chemical engineers on the next stages: cloning, creating vectors, developing bioreactors for cell growth, and purifying and checking the protein products. The industry is relying more on computer-based technologies and '-omic' methodologies to weed out toxic compounds earlier in development, including using tests with computers, called biosimulation, to cut down on testing times and identify biological markers of toxicity (see *Nature* **428**, 450; 2004).

Hitting the spot

Outsiders often overlook the importance of formulation on a drug candidate's passage through preclinical development. But without careful planning, a drug may not get to its site of action. Patch, pill or intravenous injection? Does it need to be protected until it gets to the intestine? At what pH is the drug most soluble? Will it be stable at a variety of temperatures for long periods? These are some of the questions that formulation scientists ask, to understand the physical properties of the drug, its target and intended market. They tend to be biochemists, physical chemists and scientists trained in pharmacy.

Scientific credentials remain the key to a career in industry, but hiring managers look for additional skills and experience from job applicants. "You need to be a team player, rather than an individual contributor," says Kevin Stark, senior director of strategic operations at Amgen. PhD or postdoc-level scientists should highlight any mentoring experience they've provided to young people, as they'll be expected to grow into

HIT THE LAB RUNNING

"It used to be true in the chemical industry that you could only learn on the job," says Frank Douglas, who until August 2004 was head of the division of drug innovation and approval at Aventis Pharma. Douglas is trying to change that with the recent introduction of a programme offered by the Center for Biomedical Innovation at the Massachusetts Institute of Technology.

Although young, the programme has received positive feedback. One enthusiast is David Nicholson (below) of Organon International. "The course is geared to teach students to be focused on modern drug discovery," says Nicholson. "We need more of that type of course."

Backed by more than 20 years of experience in the drug industry and the enthusiasm of industry leaders, Douglas is bringing industry's past problems into the classroom. Every two weeks, he invites a senior scientist to present a real case of a failed drug.



"The students are challenged as a team to evaluate how they would use current technologies and methodologies that may not have been available at the time when the company was dealing with these issues," says Douglas.

Next year, he hopes to add a 'biopharm academy' for promising mid-level industry scientists.

"A common experience occurs when you put a high-potential scientist into a project-leader position," he says. "Initially they struggle because they are uncomfortable that they do not have the expertise either upstream or downstream of their own particular disciplines." The course will provide them with the broad overview, an understanding of the strategic issues, the framework and tools to become leaders, Douglas says. **H.H.**

that role within the company, often managing a team of up to five bachelor's or master's level scientists.

Ingelise Saunders, chief executive of ACE BioSciences, a Danish biotech whose pipeline includes prophylactic vaccines against traveller's diarrhoea, emphasizes experience (see *Nature* **433**, 442; 2005). The company's size dictates this preference: with only ten people working in R&D, the vaccinologists she hires must be well-versed in animal models, adjuvants, immunology and data analysis, usually with a couple of years' experience at a scientific institution or in industry.

Larger companies may hire ten times more scientists with bachelor's and master's degrees than with PhDs or postdoctoral experience. For these young scientists, prior industry experience will catch the eye of any recruiter. Young scientists interested in entering drug development should look to internships, co-op placements or sandwich years to beef up their CVs (see *Nature* **439**, 504–505; 2006). It's an opportunity to gain valuable experience and to decide whether this is the work for them. "It's a good way, even as undergraduates, to get a feel for what research is like in the pharmaceutical industry," says Biller.

As R&D becomes increasingly automated, these young scientists may be asked to contribute more brain power. Hiring managers are nearly unanimous: young scientists must learn to think, or they won't be prepared for a career in industry (see box). "They know the textbook stuff, but when you get into discussions and challenges and coming up with new ideas, they haven't been used to that," says Saunders.

Nicholson stresses that to succeed in industry young scientists must be keen to discover and develop drugs to treat disease. Batchelor agrees. "In industry you are completely focused on the discovery of a pharmaceutical product that will treat patients. In academia you worry if you can publish your paper in *Nature*."

Hannah Hoag is a freelance writer based in Montreal, Canada.



The right stuff: Kevin Stark, far left, seeks team players; Ingelise Saunders looks for experienced all-rounders.

MOVERS

Thomas Baer, executive director, Stanford Photonics Research Center, California



2004–present: Consulting professor, Department of Applied Physics, Stanford University, California
1996–2005: Founder, chief executive and chairman, Arcturus Bioscience, Mountain View, California
1993–96: Vice-president, research, Biometric Imaging, Mountain View, California

As a child, Thomas Baer taught himself to build receivers and design circuits. But he credits his liberal arts education from Lawrence University in Appleton, Wisconsin, for the skills to learn disparate fields later. Staying true to his love of applied mathematics, Baer did postgraduate degrees in atomic physics at the University of Chicago. But after a postdoc with Nobel laureate John Hall at the University of Colorado, Boulder, he moved into industrial science.

Baer's entrepreneurial leanings emerged when he founded an advanced product technology group while vice-president of research at laser company Spectra-Physics. Working on a solid-state laser system, Baer invented a laser that, combined with a drug activated by light, can be used to treat cancer. He also co-founded Spectra-Physics Laser Diode Systems to develop the technology group's advanced products, including materials and applications. He would later draw on this business experience to run start-up companies.

The Human Genome Project inspired Baer to learn about biotech. "Obviously biotech was undergoing a scientific revolution that would impact society as a whole and, more specifically, my family and especially my children," he says. Without even a high-school biology course under his belt, Baer accepted an offer from a former Spectra-Physics colleague who was starting a medical diagnostic company called Biometric Imaging. "It was ridiculous to be in Silicon Valley and not become involved in a start-up venture," he says. By night, Baer studied cell-biology and immunology graduate texts, and went on to build an improved laser scanner to do white-blood-cell counts in AIDS patients. After two years of on-the-job experience, he started his own business. His company, Arcturus Bioscience, pioneered the area of microgenomics, using molecular tools to analyse single-cell samples.

Convinced that Arcturus had grown through its entrepreneurial stage and could best evolve under new leadership, he recently resigned as chief executive to become executive director of the Stanford Photonics Research Center, where he has been a member of the applied-physics faculty since 2004. The challenge at Stanford is to organize the photonics community — which at present, he says, is like a bunch of rock bands jamming — into a coordinated ensemble. But Baer acknowledges that the allure of starting high-tech companies makes it unlikely that academia will be his permanent home. ■
Virginia Gewin

RECRUITERS & ACADEMIA

Finding your north

With faculty positions dwindling and postdocs proliferating, scientists must seek other employment that uses their academic training. A daunting task, but a new website seeks to make the process easier. It provides visitors — undergraduates, postgrads, medical students, postdocs and biotech/pharmaceutical professionals — with information to help them answer the question: "What is my perfect career?"

Even though science is inherently open-minded and flexible, career advice offered to scientists-in-training is often dogmatic and restrictive. Finding Your North (FYN) seeks to provide more varied and constructive advice through articles written by professionals who have used their science background outside academia. Users can listen to interviews of people in different stages of their careers and in a variety of fields, and communicate in four groups of forums: for undergraduates, graduate students, professionals and biotech/pharmaceutical professionals.

Articles can alert users to scientific trends that may provide opportunities. One presents a model for bringing translational science to the global community, using efficacy trials of different malaria-drug combinations in African children. In another audio interview, a scientist articulates his

rationale for pursuing a job in biotech marketing in lieu of an academic career. "In my current position in marketing, I am using all my talents to make a difference in patients suffering from breast cancer," says Michael Penn, an MD/PhD and co-creator of FYN. An article giving tips on how to find a good mentor is written by a postdoc who is developing a mentorship programme for graduate students and faculty at the University of California, San Francisco.

The site's goal is to develop an exhaustive online repository of "virtual mentoring" moments that can be accessed on demand. In the future, FYN plans to stream audio and video clips of people using their science background to make a difference in society. Users will be able to see, hear and read about individuals in non-traditional scientific and medical careers. It may encourage them to carve out a unique path, if their perfect career doesn't exist yet.

FYN plans to evolve with its users' changing needs through site visitor feedback and the input of a student/postdoc advisory committee. Content will be regularly evaluated and updated. With an eye on the concerns of the next generation of scientists and physicians, FYN will strive to remain relevant. ■

Frederick Moore is co-creator of FYN.
 ▶ www.findingyournorth.com

GRADUATE JOURNAL

The great Gatsby

As an undergraduate, I received a studentship from the Gatsby Charitable Foundation that allowed me to do plant research in the United States for the summer. As a graduate student, I have just returned from a Gatsby training weekend aimed at helping PhD students to bond, learn new skills and network.

During the weekend we gave presentations recorded on video. Despite the embarrassment of seeing yourself as others see you ("Do I really have a Scottish accent?"), it made me aware that a talk is more than what you say. How your audience reacts has a lot to do with the speed at which you talk and the stance you adopt when conveying your data.

In a sense, this weekend allowed me to view my entire PhD from another perspective. Hearing about the paths prominent scientists followed to their current positions made me realize that science is a career too. What questions do I want to answer? What makes my work unique?

Their perspectives led me beyond just science to outreach. Getting children excited about science would be a great thing. And I realized that scientists need a life beyond work. I once thought that to be a great scientist you must forfeit your other interests, but I met people who draw and paint and write opera; science seems more like a hobby! Returning to the lab, I'm determined to make the most of my time as a graduate student — both in the lab and outside. ■
Mhairi Dupré is a first-year PhD student in evolutionary developmental biology at the University of Oxford, UK.

The perfect lover

In the heat of passion.

Paul Di Filippo

Neurosciences Institute, La Jolla, 10 February 2036. The substrate for the cultured human-mouse brain cells was a highly reticulated wodge of aerogel inside a homeostatic capsule as big as a thumb. At this moment the naked capsule sat in a dock, tethered by a GliaWire connection to a Brookswell 5000 running at 100 petaflops. The parent machine was the size of a credit card, its monitor and keyboard hologrammatic projections.

Two people stood by. One, a genially abstracted man of about 30, wore intelligent otakuwear, full of membranous pockets, organic sensors, interface patches and invisible circuitry. The other, a hard-eyed woman with some grey threading her bronze hair, wore the dress uniform of a Marine major, including ribbons from the Caracas campaign.

"I don't understand," said the woman, "why the drone can't be governed directly by the Brookswell. Surely there's enough Turingosity there."

"Plenty," replied the man. "But there's no love."

"Love? What's love got to do with it?"

Filtering the conversation in real time, the man's clothing prompted him through an earbud with a cultural referent to a pop song more than 50 years old. But he chose not to utter it. Didn't seem likely this hard case would appreciate any such trivial allusion. Intelligence amplification still required human discretion.

"Love is the driver for the mission. Love will supplement the drone's heuristics in instances when lesser imperatives would collapse. Without it, the failure rate goes up an order of magnitude. And we can't simulate love yet in purely moletronic minds."

The major looked suspiciously at the little pod full of wetware, as if it might suddenly start spouting poetry through as-yet-unattached peripherals.

"Well, as long as it follows its directives..."

"Need I remind you of our past successes? DARPA and BARDA just renewed our funding with double the budget."

"I know, I know. But there's so much riding on this mission. If we don't stop this bastard Kiet the Mousekiller, we stand to lose most of the West Coast."

Kiet the Mousekiller had begun his infamous career as a simple Thai pirate. Radicalized by the anonymous contamination

of Mecca with a GPS-circumscribed green goo, he had become a terrorist, earning his sobriquet by his destruction of Hong Kong Disneyland. Kiet's latest scheme, not yet known to the public, involved a retired Japanese deep-sea drilling ship, the *Chikyu*, which Kiet had purchased on the open market under a front. Now docked in an Indonesian port, the ship was believed to be due to sail imminently.

Kiet's plan was to drill into a tectonic subduction zone close to America, plant a small nuclear bomb and detonate it,



triggering a tsunami. Stopping him by overt military means was made impossible by the terrorist's current refuge with an ostensible ally. Hence the black-budget project.

After regarding the Brookswell's display, the technician began disconnecting the GliaWire. "OK, we'll be ready for the sample in a moment. You've got it?"

The major's hand strayed instinctively to her sidearm, before she reached into her pocket and removed a glassine packet. "Several hairs reclaimed from Kiet's last visit to his favourite whorehouse."

Handling the homeostatic capsule nonchalantly, the man approached the drone.

A stealth tortoise with a MEMS shell, powered by the same pocket fusion reactor as NASA's Sedna probe, the drone rested on a table, as innocuous as any lawn-mowing bot. A small hatch gaped in its shell. The technician installed the pod and closed the hatch. He took the packet, extracted the hairs, and pressed them into a small perforated depression on the front of the tortoise.

"OK, we're live."

When I came fully awake the essence of my beloved was already integrated into my soul. His beautiful face filled my inner eye, and I could taste his genome, sweeter to

me than the power that flowed from my atomic heart. I wanted nothing more than to be with him, to merge my soul with his, to shower him with my love. I would let nothing stand between us.

I immediately extended my senses, sniffing the air, but met disappointment. My beloved was nowhere within range. But knowledge in my memory informed me of his location! How I quivered with eagerness to race to his side! But where was the exit from this place? Suddenly a passage to the open air materialized above me. I activated my ventral lifter fans and rose upward.

Banda Sea, 14 February 2036. I had been damaged on my voyage to my mate. He was surrounded by vigilant duennas, brutish entities similar to myself who guarded him jealously. Every step of my route during the last day had been fraught with challenges. But I had met them without hesitation. Because that is what lovers do.

My aerial capacity was now limited to short hops. Currently, I travelled underwater, using my magneto-hydrodynamic systems. My signature across the spectrum was that of a school of fish. All my telemetry said abort: but I would not. Ahead of me loomed the vessel that held my beloved. I knew I would have to surface to unite with him, and prepared myself.

I shot out of the water alongside the ship, lurching evasively, to be met with a hail of small-arms fire from those who were not my beloved. I triggered my infrasounds, and all my rivals collapsed in bowel-spasming pain.

Crashing through the window of the pilothouse, I sustained further injury.

But nothing mattered.

For I was finally in the presence of my beloved!

An expression of terrible ecstasy filled his face, and my soul melted with joy.

I initiated the destabilizing quench on the magnets surrounding my fiery heart, giving him all my love at last.

An evanescent fountain of multi-million-degree plasma bloomed briefly aboard the *Chikyu*, in the fierce and tender shape of a heart.

A native Rhode Islander, Paul Di Filippo has lived for the past 30 years in the lovecraftian city of Providence. His latest book is *Shuteye for the Timebroker* (Thunder's Mouth Press, May). June will see the publication of his graphic novel *Top 10: Beyond the Farthest Precinct*.

JACEY